

THE EFFECTS OF DIFFERENTIAL ITEM FUNCTIONING ON PREDICTIVE BIAS

by

DAMON ULYSSES BRYANT  
B.S. Howard University, 1995  
M.A.G.P. University of North Florida, 2000

A dissertation submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy  
in the Department of Psychology  
in the College of Arts & Sciences  
at the University of Central Florida  
Orlando, Florida

Fall Term  
2004

Major Professor: Dr. Eugene Stone-Romero

## ABSTRACT

The purpose of this research was to investigate the relation between measurement bias at the item level (differential item functioning, DIF) and predictive bias at the test score level. DIF was defined as a difference in the probability of getting a test item correct for examinees with the same ability but from different subgroups. Predictive bias was defined as a difference in subgroup regression intercepts and/or slopes in predicting a criterion. Data were simulated by computer. Two hypothetical subgroups (a reference group and a focal group) were used. The predictor was a composite score on a dimensionally complex test with 60 items. Sample size (35, 70, and 105 per group), validity coefficient ( $\rho = .3$  or  $.5$ ), and the mean difference on the predictor (0,  $.33$ ,  $.66$ , and 1 standard deviation, *SD*) and the criterion (0 and  $.35$  *SD*) were manipulated. The percentage of items showing DIF (0%, 15%, and 30%) and the effect size of DIF (small =  $.3$ , medium =  $.6$ , and large =  $.9$ ) were also manipulated. Each of the 432 conditions in the  $3 \times 2 \times 4 \times 2 \times 3 \times 3$  design was replicated 500 times. For each replication, a predictive bias analysis was conducted, and the detection of predictive bias against each subgroup was the dependent variable. The percentage of DIF and the effect size of DIF were hypothesized to influence the detection of predictive bias; hypotheses were also advanced about the influence of sample size and mean subgroup differences on the predictor and criterion. Results indicated that DIF was not related to the probability of detecting predictive bias against any subgroup. Results were inconsistent with the notion that measurement bias and predictive bias are mutually supportive, i.e., the presence (or absence) of one type of bias is evidence in support of the presence (or absence) of the other type of bias. Sample size and mean differences on the

predictor/criterion had direct and indirect effects on the probability of detecting predictive bias against both reference and focal groups. Implications for future research are discussed.

I dedicate this work to Mr. and Mrs. Harold Damon Bryant, Sr.

## ACKNOWLEDGMENTS

I would like to thank all the members of my Dissertation Committee: Dr. Eugene Stone-Romero, Dr. William Wooten, Dr. Jian-Jian Ren, and Dr. Dianna Stone. I am grateful to all of the faculty and staff associated with the Industrial and Organizational Psychology Ph.D. Program for providing an environment that was conducive to learning. I would also like to thank the Florida Education Fund and the Educational Testing Service for providing the resources necessary to complete this work. This research was sponsored in part by a Harold Gulliksen Psychometric Fellowship awarded by the Educational Testing Service.

## TABLE OF CONTENTS

LIST OF FIGURES .....	viii
LIST OF TABLES .....	x
LIST OF ABBREVIATIONS .....	xii
INTRODUCTION .....	1
Purpose and Research Questions .....	2
Validity in Cognitive Ability Testing .....	3
Forms of Validity Evidence .....	3
Fairness in Cognitive Ability Testing .....	8
Differential Item Functioning .....	9
Predictive Bias .....	25
Link Between DIF and Predictive Bias .....	35
Hypotheses .....	39
METHOD .....	43
Fixed Conditions .....	43
Test Design .....	44
Independent Variables .....	48
Dependent Variable .....	54
Data Analysis .....	54
RESULTS .....	55
Summary Tables .....	56
Reference Group Analysis .....	72

Focal Group Analysis .....	77
Exploratory Analyses.....	80
DISCUSSION.....	89
Summary of Findings.....	89
Conclusions.....	97
Scientific and Social Implications .....	100
APPENDIX A: PROOF OF THETA MAXIMUM AND ITEM INFORMATION .....	105
The Item Response Model .....	106
Item Information Function.....	107
Theta Maximum for the Multidimensional 3-PL Model .....	109
APPENDIX B: ALGORITHM FOR THE SIMULATION STUDY .....	116
REFERENCES .....	122

## LIST OF FIGURES

Figure 1. The 1-PL model with $b_i = 1$ .....	13
Figure 2. The 2-PL model with $b_i = 1$ and $a_i = 1.5$ .....	14
Figure 3. The 3-PL model with $b_i = 1$ , $a_i = 1.5$ , and $c_i = .25$ .....	15
Figure 4. Uniform DIF: Reference group ( $b_i = .8$ , $a_i = 1$ ) and focal group ( $b_i = 1.3$ , $a_i = 1$ ).....	21
Figure 5. Non-Uniform DIF: Reference group ( $b_i = 0$ , $a_i = 1$ ) and focal group ( $b_i = 0$ , $a_i = .6$ )... ..	22
Figure 6. Predictive bias with different subgroup intercepts and equal slopes.....	31
Figure 7. Predictive bias with different subgroup slopes.....	33
Figure 8. Equality of subgroup intercepts and slopes.....	33
Figure 9. Geometric representation of items in two dimensions.....	46
Figure 10. Test information for the theta composite, $\theta_c$ .....	47
Figure 11. Standard error of estimation for the theta composite, $\theta_c$ .....	47
Figure 12. Predictive bias as a function of sample size: Reference group.....	75
Figure 13. Predictive bias as a function of criterion difference: Reference group.....	76
Figure 14. Predictive bias as a function of sample size: Focal group.....	78
Figure 15. Predictive bias as a function of predictor difference: Focal group.....	79
Figure 16. Predictive bias as a function of predictor and criterion difference: Reference group.....	82
Figure 17. Predictive bias as a function of predictor and criterion difference: Focal group.....	85
Figure 18. Predictive bias as a function of predictor and criterion difference when percentage of DIF = 0: Focal group.....	86

Figure 19. Predictive bias as a function of predictor and criterion difference when percentage of DIF = 15: Focal group. ....	87
Figure 20. Predictive bias as a function of predictor and criterion difference when percentage of DIF = 30: Focal group. ....	88
Figure 21. Diagram of the object-oriented program architecture. ....	121

## LIST OF TABLES

Table 1. Unidimensional 1-, 2-, and 3-Parameter Logistic Models.....	18
Table 2. Multidimensional 1-, 2-, and 3-Parameter Logistic Models.....	19
Table 3. Item Parameters for the Multidimensional 2-Parameter Logistic Model .....	49
Table 4. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and No Differential Item Functioning (% of DIF = 0, Effect Size of DIF = 0) .....	58
Table 5. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor and Criterion, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 15, Effect Size of DIF = .3) .....	60
Table 6. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 15, Effect Size of DIF = .6) .....	62
Table 7. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 15, Effect Size of DIF = .9) .....	64
Table 8. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the	

Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 30, Effect Size of DIF = .3) .....	66
Table 9. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 30, Effect Size of DIF = .6) .....	68
Table 10. Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 30, Effect Size of DIF = .9) .....	70
Table 11. Summary of the Logistic Regression Analysis for Variables Predicting Predictive Bias Against the Reference Group ( $N = 864$ ) .....	73
Table 12. Summary of the Logistic Regression Analysis for Variables Predicting Predictive Bias Against the Focal Group ( $N = 864$ ).....	77

## LIST OF ABBREVIATIONS

CTT	Classical Test Theory
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
IIF	Item Information Function
IRT	Item Response Theory
M2PL	Multidimensional 2-Parameter Logistic
M3PL	Multidimensional 3-Parameter Logistic
MDISC	Multidimensional Discrimination
MDIFF	Multidimensional Difficulty
PL	Parameter Logistic
TIF	Test Information Function

## INTRODUCTION

Cognitive ability testing is one of the most widely used methods for deciding who will receive favorable or unfavorable outcomes in educational and employment settings in the United States and abroad (Huysamen, 2002; Muchinsky, 1993; Sackett, Schmitt, Ellingson, & Kabin, 2001; Vincent, 1996). Because of the ubiquitous nature of cognitive ability testing and the large number of people wanting favorable outcomes (e.g., a high test score), decision makers readily act on results of tests as an efficient means for the distribution of such limited resources as jobs, licenses, and class seats (Cleary, Humphreys, Kendrick, & Wesman, 1975; Florida Department of Education, 2002). Consequently, there may be a perception of unfairness or bias when outcomes for one person or group are not equal to outcomes of another person or group of comparable standing (Adams, 1963). This perception of unfairness may lead to costly litigation and test legislation (McAllister, 1993).

Professional testing guidelines suggest that both item bias and predictive bias studies be done to ensure that the inferences made from test score interpretations are fair to persons who vary in terms of such variables as sex and ethnic background (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society for Industrial and Organizational Psychology, 2003). Despite the recommendation of these guidelines to investigate bias whenever feasible, item bias and predictive bias have substantially different research foundations and often answer very different questions related to the validity of test score interpretations (Camilli & Shepard, 1994; Osterlind, 1980). Because there are different research foundations for the investigation of item and predictive bias, it is possible that ambiguity about the relation between these two areas of test

fairness research exists (Hunter & Schmidt, 2000; Hunter, Schmidt, & Rauschenberger, 1984; Jensen, 1980; Millsap, 1997).

### *Purpose and Research Questions*

The purpose of this study was to investigate the relation between measurement bias at the item level and predictive bias at the test score level. An important question this study sought to answer is the following: What effect does measurement bias at the item level have on the detection of predictive bias at the test score level? Some researchers posit that measurement bias and predictive bias are mutually supportive (Hunter & Schmidt, 2000; Hunter et al., 1984; Jensen, 1980). In other words, the presence or absence of one type of bias is evidence in support of the presence or absence of the other type of bias. For example, Hunter and Schmidt (2000) have argued recently that because the literature shows no evidence of predictive bias against minority groups, then item bias or differential item functioning (DIF) does not exist against minority groups, which implies that the lack of predictive bias should be taken as evidence of no item bias or DIF. This leads to two secondary questions about the relation between predictive bias and DIF. Can predictive bias against a subgroup exist when DIF is not present? Can DIF against a subgroup exist when predictive bias is not present?

The outline is as follows. First, validity and fairness are reviewed in relation to the testing literature. The review is based on the guidelines that are currently being used to develop cognitive ability tests in education and employment. Second, theory and methods for investigating item and predictive bias in cognitive ability tests are briefly presented. Third, a presentation of the similarities and differences between measurement bias and predictive bias is

given. Fourth, hypotheses are advanced about the relation between measurement bias at the item level and predictive bias at the test score level.

### *Validity in Cognitive Ability Testing*

The American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have created guidelines for developing and using cognitive ability tests. The same guidelines, used primarily for the development of tests in education and employment, are the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999; APA, 1966, 1974), hereafter referred to as the *Standards*. Guidelines developed by the Society for Industrial and Organizational Psychology (SIOP) for creating cognitive ability tests in employment contexts are the *Principles for the Validation and Use of Personnel Selection Procedures* (APA, 1980; SIOP, 2003), hereafter referred to as the *Principles*. The *Principles* have been revised to be consistent with the most recent version of the *Standards* (SIOP, 2003). Although there is no enforcement authority for these guidelines, the *Standards* and *Principles* both represent a consensus of practices, which are used as prescriptions for the design, validation, and use of cognitive ability tests and other assessments.

### *Forms of Validity Evidence*

One of the most critical aspects in testing is the validity of test score interpretations (AERA, APA, & NCME, 1999; APA, 1966, 1974, 1980; Cronbach & Meehl, 1959; Messick, 1980, 1989, 1995a, 1995b; SIOP, 2003). Validity has evolved over the years from a view based

on three separate components, i.e., content, construct, and criterion-related validity, to a view based on a unitary concept (Messick, 1989, 1995b). According to the *Standards*, “[v]alidity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, APA, & NCME, 1999, p. 9). In other words, test scores, interpretations of test scores, and the inferences made from the interpretations must be justified by theory and empirical research evidence.

More recent conceptions of validity pertain to various sources of evidence collected to shore up test score interpretation and use (Messick, 1989, 1995a, 1995b). These forms of evidence include those based on test content (Messick, 1989), substantive aspects of the construct domain (Embretson, 1983; Embretson & Reise, 2000; Kyllonen & Christal, 1989), internal structure of the test (Loevinger, 1957; Reckase, 1997), generalizability of test score meaning (Cook & Campbell, 1979), relations with external criteria (Campbell & Fiske, 1959) and social consequences of testing (Messick, 1995a, 1995b). Each is briefly summarized below.

*Test content.* An important part of any construct validation effort relevant to content involves delineating the boundaries of the construct domain to be assessed (AERA, APA, & NCME, 1999). The forms of evidence appropriate for the content aspect of validity deal with relevance, representativeness, and quality of what is being used as a measure of the focal construct (Messick, 1995b). In employment, this evidence can be obtained by documenting information from an interview with a subject matter expert about particular duties required on the job or by perusing results from job and task analyses (SIOP, 2003). In education, a curriculum analysis and expert judgment may provide cogent evidence of the appropriateness and relevance of the content of a test (Messick, 1989).

*Substantive aspect of the construct domain.* The substantive aspect of construct validation has its basis in cognitive theories and process models of task performance, which are used to describe the mental activities purported to occur during task performance (AERA, APA, & NCME, 1999; Embretson, 1983, 1995, 2002; Kyllonen & Christal, 1989; Messick, 1996). Forms of evidence representative of this aspect of construct validation include “think aloud” protocols, evidence of performance consistent with a cognitive model (Kyllonen & Christal, 1989), and the analysis of residual outcomes in computer-based assessments, e.g., response time (Embretson, 1995). Investigations of this type give insight into ancillary tasks or processes, which may lead to different consequences for different subgroups.

*Internal structure of the test.* An investigation of the structure of a test provides support for the notion that its internal structure corresponds to the internal structure of the construct domain to be measured (Loevinger, 1957; Messick, 1995a; Reckase, 1997). Moreover, research should investigate the internal structure to determine if items within a test function differently for various subpopulations (AERA, APA, & NCME, 1999; Hambleton, Swaminathan, & Rogers, 1991; Mellenbergh, 1982; Roussos & Stout, 1996; Rudner, Getson, & Knight, 1980); these are known as differential item functioning or DIF analyses (Mellenbergh, 1982). Camilli (1993) argues that DIF studies address questions of construct representation, which centers on the internal structure of a test. Other forms of evidence about the internal structure might include studies of inter-item relations and relations among subcomponents of the construct domain (Messick, 1989). For the purpose of this study, emphasis is placed on DIF for assessing the internal structure of a test.

*Generalizability of test score meaning.* Inquiries about the degree to which score interpretations can be extended to other settings may provide evidence about the generalizability

of score meaning (Cook & Campbell, 1979; Messick, 1989, 1995b). If results from a study can be readily generalized to other people, places, settings, and times, the meaning of scores may add to the overall validity of test score interpretations. However, it should be noted that instruments designed in one setting may not necessarily have the same interpretation in a different setting (Cole & Moss, 1989). For example, a criterion-referenced test designed to measure if a student has obtained a certain level of achievement on a dimension of interest may not be pertinent in understanding the same student's standing relative to others in some population. In addition, a high-stakes exam used for promotion to police sergeant in a precinct may be totally inappropriate for selecting police recruits for a training academy. Thus, a consistent meaning of test score interpretation in different places, settings, and times is an important aspect of validity.

*Relations with external criteria.* Forms of evidence that have a bearing on the external aspects of construct validation include studies delineating convergent and discriminant relations among the focal construct intended to be measured by the test and the indicators of other variables in the nomological network (Campbell & Fiske, 1959). Specifically, these forms of evidence elucidate the relation that test scores have with a criterion for which the test was developed (Cleary, 1968). Investigations of this type include predictive and concurrent validation studies (AERA, APA, & NCME, 1999; Cascio, 1998). These studies can be used to examine the degree to which score meanings and the relations of scores to external criteria differ across subgroups of interest, perhaps due to construct under-representation or construct-irrelevant variance of the test (Embretson, 1983; Messick, 1989).

*Social consequences of testing.* Evidence pertaining to the intended and unintended social consequences of testing should be documented (AERA, APA, & NCME, 1999). Long-term and

short-term uses of tests are investigated to determine the potential for bias in scoring, interpretation, and inferences made from test scores, which might result in unfair test use (Messick, 1995). For example, although the *Standards* recommend that a single test score should not be used to make a high-stakes decision, the Florida Comprehensive Assessment Test is utilized to determine if elementary school students will be promoted to the fourth grade; it is also used to award diplomas to high school seniors, even though the psychometric properties of the test are questionable (Florida Department of Education, 2002). In this respect, social values play a very important role in deciding whether to employ a test for a specific purpose (Messick, 1980).

The validity of test score interpretation is a judgment reached by evaluating the aforementioned forms of evidence used to buttress the intended uses of a test. Construct validity of the test is threatened when an evaluation of the integrative summary of evidence is inconsistent with the proposed uses of the test. Sources of invalidity can be due to construct under-representation and construct-irrelevant variance (Messick, 1995b). Construct under-representation is the degree to which the test measuring the construct is too narrow and does not include important dimensions specified in the construct domain. Construct-irrelevant variance is a reliable component of test scores that is outside of the bounds of the construct domain. For example, if a test is designed to measure only algebraic reasoning, but items on the test rely heavily on the ability to analyze and extract information from graphs, then graph-reading ability may be considered a construct-irrelevant aspect of the test. As in the above example, performance on the test may involve irrelevant abilities that contribute to systematic test score differences (Embretson, 1983; Messick, 1995b). Test score interpretations and inferences may be

inaccurate, which may lead to different outcomes for different subgroups and foster perceptions of unfairness or inequity (Adams, 1963).

### *Fairness in Cognitive Ability Testing*

According to the most recent version of the *Standards and Principles*, fairness has several meanings that include the following: (a) lack of bias, (b) equitable treatment, (c) equality of outcomes for all, and (d) opportunity to learn (AERA, APA, & NCME, 1999; SIOP, 2003). Although some of these meanings have been rejected by the *Standards* (e.g., equality of outcomes for all), other meanings are considered seriously. However, for the purpose of this study, emphasis is placed on fairness as a lack of bias. The rationale for this focus is as follows. All other aspects of fairness mentioned by the *Standards* imply that the instrument used to measure the construct yields unbiased numerical scores, which are then subject to interpretations and inferences. If numerical scores are systematically biased, the interpretation and meaning of test scores are untenable (AERA, APA, & NCME, 1999). Thus, a lack of bias in test scores is central to fairness.

The *Standards and Principles* identify two types of bias investigations: item bias and predictive bias (AERA, APA, & NCME, 1999; SIOP, 2003). As stated above in the section on validity, investigations of item bias (i.e., DIF) provide evidence of a consistent internal structure of the test across subpopulations and do not require an external criterion (Hambleton et al., 1991; Osterlind, 1983; Raju, 1988; Raju & Ellis, 2002). In contrast, predictive bias studies give information about the relation between the construct (i.e., the predictor as measured) and an external criterion of interest, which is assumed to be free of bias (Lewis-Beck, 1980; Schmidt &

Hunter, 1974, 1984). Even though there are noticeable similarities and differences between these two types of bias investigations in terms of the questions they seek to answer (Jones & Applebaum, 1989), each serves to provide important evidence in support of the validity of test score interpretation and use (AERA, APA, & NCME, 1999). An overview of each type of bias follows, which includes the theoretical background, a definition, and the empirical evidence relevant to each type of bias.

### *Differential Item Functioning*

*Background.* Item bias studies are concerned with measurement bias at the item level. These studies are also known as DIF analyses. DIF has advanced under the auspices of Item Response Theory (IRT; Hambleton et al., 1991; Lord, 1980; Lord & Novick, 1968). Information on the theoretical framework (IRT) that makes the investigation of DIF feasible is first discussed. Next a short overview on the history of DIF investigations is given, which is followed by a definition of DIF. Some methods used to detect DIF are briefly surveyed. Then, evidence of DIF in the domain of cognitive ability testing is presented.

IRT provides an adequate framework for the investigation of DIF (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Lord, 1980; Nunnally & Bernstein, 1994). Some advantages of IRT over Classical Test Theory (CTT; Gulliksen, 1950) are in areas such as test equating and computer-adaptive testing (Hambleton et al., 1991). Although CTT is still used today and is often linked to IRT (Dimitrov, 2003), a presentation of the relative advantages and disadvantages of IRT versus CTT is not offered here (for a thorough comparison of CTT and IRT, see Embretson & Reise, 2000). Now recognized as modern test theory (Nunnally &

Bernstein, 1994), IRT makes several important assumptions: monotonicity, local independence, and unidimensionality. Each is discussed in turn.

*Monotonicity.* The first assumption is monotonicity, which posits that as ability increases, the chance or probability of getting an item correct increases. In other words, the function is non-decreasing with an increase in the ability or latent trait ( $\theta$ ) being measured by the test (Lord & Novick, 1968). Because most responses to items are scored correct (1) or incorrect (0), the form of the relation between  $\theta$  and the probability of a correct response is not linear but S-shaped. This S-shaped curve is called an item characteristic curve (ICC; Lord & Novick, 1968).

*Local independence.* Another assumption of IRT is local independence, which states that the relation between any two items is independent while conditioning on a local or specific point on the latent trait continuum (Hambleton & Swaminathan, 1985). In other words, when  $\theta$  is held constant, the conditional distributions of responses to different items should be orthogonal (Embretson & Reise, 2000; Hambleton et al., 1991; Lord 1980; Lord & Novick, 1968); a parallel to this can be drawn to the CTT notion of independence of error of two tests once true score variance is taken into account (Gulliksen, 1950). When the assumption of local independence holds, the probability or likelihood of any sequence of item responses occurring is simply the product of all individual item probabilities (Hambleton et al., 1991); this has important implications for the estimation of item parameters and the investigation of DIF (McDonald & Mok, 1995).

*Unidimensionality.* The third assumption is unidimensionality, which means that responses to test items are a function of one underlying dimension (Lord, 1980). If responses on the test require only one trait or ability in getting the answer correct, then unidimensionality is

satisfied. However, if more than one ability is required in getting an item right, and there is a between-group difference in the conditional distributions of one ability while holding another ability constant, then major problems occur when a unidimensional measurement model is used. Not only is the assumption of unidimensionality violated, but also local independence is untenable (Junker, 1982).

The assumption of unidimensionality has recently come under sharp criticism by some psychometricians (Hunter & Schmidt, 2000; Miller & Hirsch, 1992; Reckase, 1985; Reckase & McKinley, 1991; Roussos & Stout, 1996; Stout, 1990) and has been relaxed to be either essentially unidimensional (Stout, 1990) or intentionally multidimensional (Reckase, 1997). Multidimensional measurement models provide a way of reinstating local independence by accommodating multiple ability estimates (Ackerman, 1994b). Not only do these models allow for the estimation of multiple traits and standard errors associated with each trait estimate, but they also permit computation of a variety of composite scores with measurements of quality for each composite (Ackerman, 1994b). This move to consider tests as having either one dominant dimension (essentially unidimensional) or multiple dimensions (intentionally multidimensional) suggests that one test score can be dimensionally complex; it also provides a means for thoroughly understanding what a test is measuring (Ackerman, 1994a). As stated in the *Principles*, most tests used in employment settings measure more than one relevant dimension and may not meet the assumption of unidimensionality (SIOP, 2003); this position also has been promulgated by other researchers (Hunter & Schmidt, 2000). Thus, multidimensional measurement models may be a more feasible alternative for tests that involve the use of multiple skills or abilities (Reckase, 1985, 1997; Reckase & McKinley, 1991). In summary, the just-noted IRT assumptions provide a basis for investigating DIF.

Two functions are used to model the interaction of an examinee and an item in IRT (Birnbaum, 1968; Lord & Novick, 1968). Although the cumulative normal function was initially proposed to operationalize the ICC (Lord, 1980), Birnbaum (1968) proposed the logistic function as more tractable in estimating the curve. With the addition of a scaling constant ( $D = 1.7$ ), the difference in probabilities between the two functions is indistinguishable to the eye and is less than .01 across all levels of  $\theta$  (Birnbaum, 1968; Hambleton & Swaminathan, 1985; Mood, Graybill, & Boes, 1974). Due to the widespread use of the logistic function in high-stakes testing (Florida Department of Education, 2002; Hambleton et al., 1991; Raju, 1990; Reckase, 1985, 1997), it is used in this study to model item responses.

*IRT parameters and models.* Each IRT logistic model has parameters that describe the form of the ICC (Lord & Novick, 1968). When additional item parameters are incorporated into the model to describe the data, larger sample sizes are also needed to estimate these parameters accurately (Ellis & Mead, 2002; Hambleton & Swaminathan, 1985). Although 200 observations have been recommended as a minimum sample size for some programs (e.g., BILOG) to estimate item parameters (Ellis & Mead, 2002), other researchers suggest that stable IRT parameter estimates can be obtained using sample sizes as small as 100 observations (Maydeu-Olivares, 2001). IRT models can have 1, 2, or 3 parameters describing aspects of the item.

The 1-Parameter Logistic (1-PL) model has a single parameter that describes the location on  $\theta$  where an examinee will have a 50% chance of getting the item correct (Hambleton et al., 1991). This is known as the median in some mathematical disciplines but is recognized as the difficulty parameter ( $b_i$ ) in IRT. For the hypothetical item described by the ICC of Figure 1, the location on  $\theta$  where examinees have a 50% chance of getting the item right is 1.0.

The 2-PL model has two parameters: one representing the item difficulty and the other representing the slope. The slope parameter is known as the discrimination parameter, which is symbolized as  $a_i$ . It gives an indication of how the item distinguishes between examinees of different  $\theta$ s. In Figure 2, the discrimination of the ICC is 1.5, and difficulty is 1.0, which means that the item is moderately difficult and quite able to distinguish high and low performer on  $\theta$ .

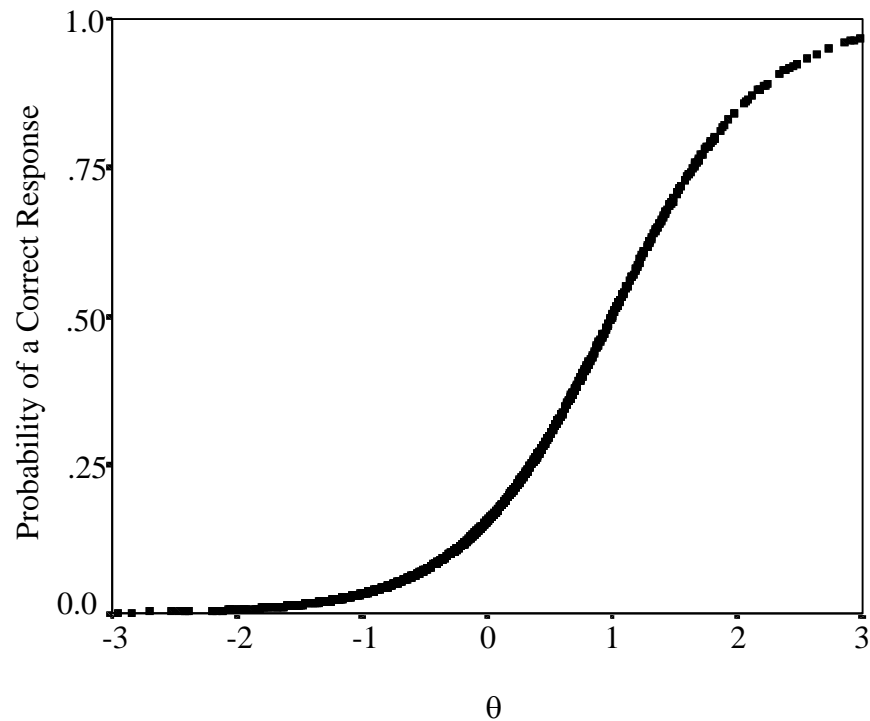


Figure 1. The 1-PL model with  $b_i = 1$ .

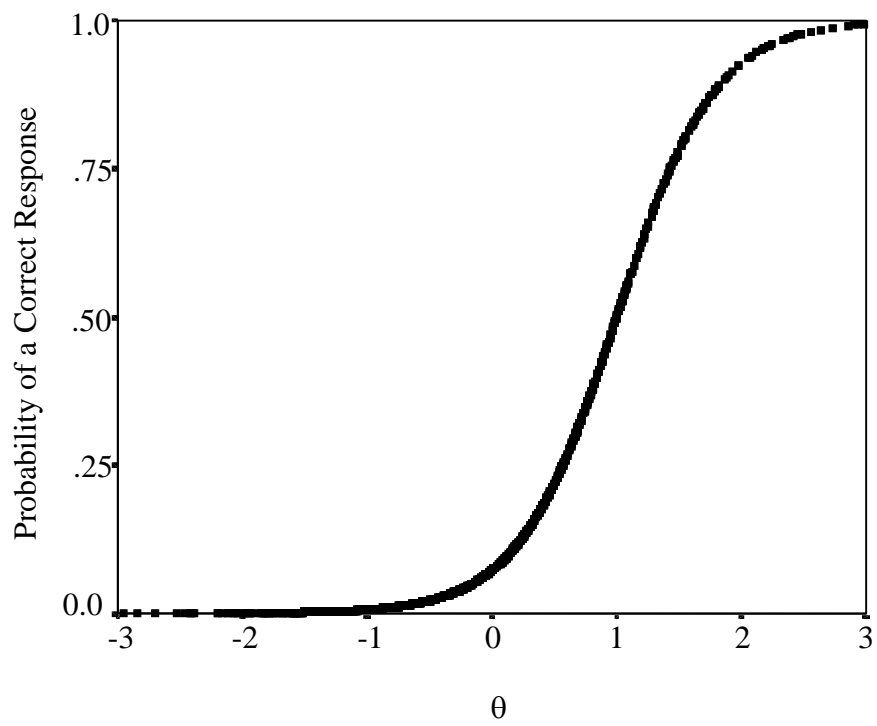


Figure 2. The 2-PL model with  $b_i = 1$  and  $a_i = 1.5$ .

In addition to discrimination and difficulty parameters, the 3-PL model has a parameter that indicates the degree to which examinees with low ability have a chance of guessing the correct answer to a question (Birnbaum, 1968; Lord, 1980). This parameter is called the pseudo-guessing parameter,  $c_i$ . As can be seen in the example shown in Figure 3, the lower asymptote is .25, which is the probability of an individual with a very low level of  $\theta$  getting the item correct. This parameter is appropriate when modeling multiple-choice responses.

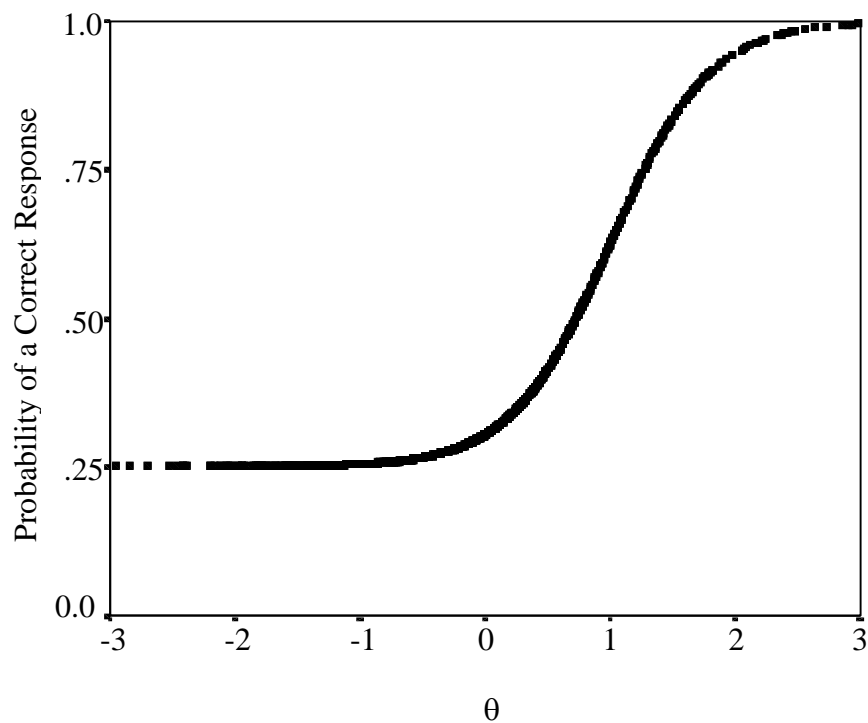


Figure 3. The 3-PL model with  $b_i = 1$ ,  $a_i = 1.5$ , and  $c_i = .25$ .

The 1-, 2-, and 3-PL models can be used to estimate one or more  $\theta$ s. Models are unidimensional if only one  $\theta$  is estimated for a person,  $j$  (Lord, 1980). If more than one  $\theta$  is measured, then the model is multidimensional. The  $\theta$ s to be estimated for person  $j$  are now represented by a vector,  $\boldsymbol{\theta}_j = [\theta_{1j}, \theta_{2j}, \dots, \theta_{kj}]'$ , where  $k$  is the number of dimensions (Reckase, 1985, 1997).

The precision or quality of an item and test can be estimated by an information function (Birnbaum, 1968; Samejima, 1977); the *Standards* recommend that item and test information be documented, especially when the psychometric quality of the item or test is being judged (AERA, APA, & NCME, 1999). Unlike in CTT, where the reliability estimate of scores is sample-dependent and is assumed to be constant across all levels of total test scores (Nunnally & Bernstein, 1994), the precision of a test in IRT is conditional on an estimate of  $\theta$  and is independent of the sample used to estimate it (Samejima, 1977). This estimate of test precision is called the Test Information Function (TIF). Because test information is inversely related to the variance of  $\theta$  (Lord, 1980), it is also used to construct confidence intervals around the  $\theta$  estimate (Hambleton & Swaminathan, 1985).

The precision of the item is computed with the Item Information Function (IIF, Birnbaum, 1968). It plays a critical role in modern test development (Embretson & Reise, 2000; Hambleton et al., 1991; Lord, 1980; Lord & Novick, 1968; Samejima, 1977). Not only is it used to design criterion-referenced and norm-referenced tests (Lord, 1980), but it is also used to compute the point of maximum item information, which is readily employed in the development and operation of computer-adaptive tests (Weiss, 1995). Because the sum of all the individual

IIFs is the total TIF, item level information plays a very important role in the design of modern tests (Samejima, 1977).

The conditional IIFs and the points of maximum item information are well known for unidimensional IRT models (Embretson & Reise, 2003; Hambleton & Swaminathan, 1985). Table 1 shows the relevant functions. However, the conditional IIFs and points of maximum item information are relatively unknown for multidimensional models, especially composites measuring more than one ability, as in employment tests (SIOP, 2003). The potential use of IRT in employment may depend on the degree to which multidimensional IRT functions are readily available for researchers and practitioners. Thus, conditional IIFs and the point of maximum item information for composites are derived for the multidimensional 3-PL model (see Appendix A). The results of the derivations in Appendix A are displayed in Table 2. The functions in the tables provide the necessary tools for developing tests and investigating measurement bias in data that are either unidimensional or multidimensional. For the purpose of this study, the multidimensional 2-PL model was used to design the dimensionally complex test.

Since the first investigations of measurement bias by Binet in 1910 and Eells in the early 1950s (Camilli & Shepard, 1994; Eells, Davis, Havighurst, Herrick, & Tyler, 1951), many definitions of item bias and methods of detecting it have been presented and were found to be unsatisfactory on one or more legitimate grounds. See Osterlind (1983) and Camilli and Shepard (1994) for a review and summary of methods and criticisms.

Table 1.

## Unidimensional 1-, 2-, and 3-Parameter Logistic Models

Model	
<u>1-PL</u>	
Item response function	$P_i(\theta_j) = \{1 + \text{Exp}[-D(\theta_j - b_i)]\}^{-1}$
Item information	$I_i(\theta) = D^2 P_i(\theta) Q_i(\theta)$
Theta maximum	$\theta_{\max} = b_i$
<u>2-PL</u>	
Item response function	$P_i(\theta_j) = \{1 + \text{Exp}[-Da_i(\theta - b_i)]\}^{-1}$
Item information	$I_i(\theta) = D^2 a_i^2 P_i(\theta) Q_i(\theta)$
Theta maximum	$\theta_{\max} = b_i$
<u>3-PL</u>	
Item response function	$P_i(\theta_j) = c_i + (1 - c_i) \{1 + \text{Exp}[-Da_i(\theta - b_i)]\}^{-1}$
Item information	$I_i(\theta) = D^2 a_i^2 Q_i(\theta) \{P_i(\theta) [1 + \text{Exp}(-L)]^2\}^{-1}$
Theta maximum	$\theta_{\max} = \ln\{.5[1 + (1 + 8c_i)^{1/2}]\} (Da_i)^{-1} + b_i$

Note.  $D = 1.7$ ,  $L = Da_i(\theta - b_i)$ , and  $Q_i(\theta) = 1 - P_i(\theta)$ .

Table 2.

## Multidimensional 1-, 2-, and 3-Parameter Logistic Models

Model	
<u>M1-PL</u>	
Item response function	$P_i(\theta_j) = \{1 + \text{Exp}[-D(\mathbf{1}'\theta_j + d_i)]\}^{-1}$
Item information	$I_{uu}(\boldsymbol{\theta}) = D^2 k P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta})$
Theta maximum	$\boldsymbol{\theta}_{\max} = [MDIFF_i / (k)^{1/2}, \dots, MDIFF_i / (k)^{1/2}]'$
<u>M2-PL</u>	
Item response function	$P_i(\boldsymbol{\theta}_j) = \{1 + \text{Exp}[-D(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)]\}^{-1}$
Item information	$I_{uu}(\boldsymbol{\theta}) = D^2 (\mathbf{a}_i' \mathbf{u}_i)^2 P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta})$
Theta maximum	$\boldsymbol{\theta}_{\max} = [MDIFF_i \cos \alpha_{1i}, \dots, MDIFF_i \cos \alpha_{ki}]'$
<u>M3-PL</u>	
Item response function	$P_i(\boldsymbol{\theta}_j) = c_i + (1 - c_i) \{1 + \text{Exp}[-D(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)]\}^{-1}$
Item information	$I_{uu}(\boldsymbol{\theta}) = D^2 (\mathbf{a}_i' \mathbf{u}_i)^2 Q_i(\boldsymbol{\theta}) \{P_i(\boldsymbol{\theta}) [1 + \text{Exp}(-L)]\}^{-1}$
Theta maximum	$\boldsymbol{\theta}_{\max} = [\ln\{.5 [1 + (8c_i + 1)^{1/2}]\} (D \cdot MDISC_i)^{-1} + MDIFF_i] \mathbf{u}_i$

Note.  $k$  = number of dimensions,  $\mathbf{1}$  is a  $k \times 1$  vector of ones.  $D = 1.7$ ,  $\mathbf{a}_i$  is a  $k \times 1$  vector of discrimination parameters for item  $i$ ,  $[a_{1i}, \dots, a_{ki}]'$ ,  $\boldsymbol{\theta}$  is a vector of  $k$  ability parameters,  $[\theta_{1j}, \dots, \theta_{kj}]'$ ,  $d_i$  is a scalar related to difficulty,  $L = D(\mathbf{a}_i' \boldsymbol{\theta} + d_i)$ ,  $MDISC_i = \|\mathbf{a}_i\|$ ,  $MDIFF_i = -d_i / \|\mathbf{a}_i\|$ , and  $\mathbf{u}$  is a vector of directional cosines,  $\mathbf{a}_i / \|\mathbf{a}_i\|$  or  $[\cos \alpha_{1i}, \dots, \cos \alpha_{ki}]'$ .

Many of the definitions of bias and methods used to detect it either failed to consider true mean differences between subgroups on the relevant ability being measured by the test or failed to appraise thoroughly the characteristics of the item that lead to the perception of item bias, such as between-item differences in difficulties (e.g., Angoff & Ford, 1973; Cleary & Hilton, 1968). CTT's indices of psychometric quality – i.e., difficulty (proportion correct,  $p_i$ ) and discrimination (point bi-serial or item-total correlation,  $r_{pb}$ ) – are sample specific and are known to confound characteristics of items and persons. It appears that IRT approaches have produced more acceptable definitions and methods of detecting item bias (Holland & Wainer, 1993).

*Definition of DIF.* Due to the necessity of a distinction between social and scientific connotations of bias, measurement bias at the item level has come to be recognized as DIF (Hambleton et al., 1991). DIF is used to describe items on a test that function differently “for two or more groups if the probability of a correct answer to a test is associated with group membership for examinees of comparable ability” (Camilli, 1993, pp. 397-398). Hambleton et al. (1991) explain the reasoning behind the use of the term DIF instead of item bias:

Investigations of bias involve empirical evidence concerning the relative performances on the test item of members of the minority group of interest and members of the group that represents the majority. Empirical evidence of differential performance is necessary, but not sufficient, to draw the conclusion that bias is present; this conclusion involves an inference that goes beyond the data. To distinguish the empirical evidence from the conclusion, the term differential item functioning (DIF) rather than bias is

commonly used to describe the empirical evidence obtained in the investigation of bias. (p. 109)

*Types of DIF.* There are two types of DIF: uniform and non-uniform DIF. Uniform DIF is described as a consistent difference in item difficulty across levels of  $\theta$  (Mellenbergh, 1982). Figure 4 shows ICCs for two subgroups: one for the reference group (e.g., Anglo-American or male) and another for the focal group (e.g., African-American or female). In the figure, the ICCs are parallel and do not intersect. To the extent that DIF of this type occurs on one or more items, shifts in subgroup means may occur, due to the accumulation of systematic error in item response curves.

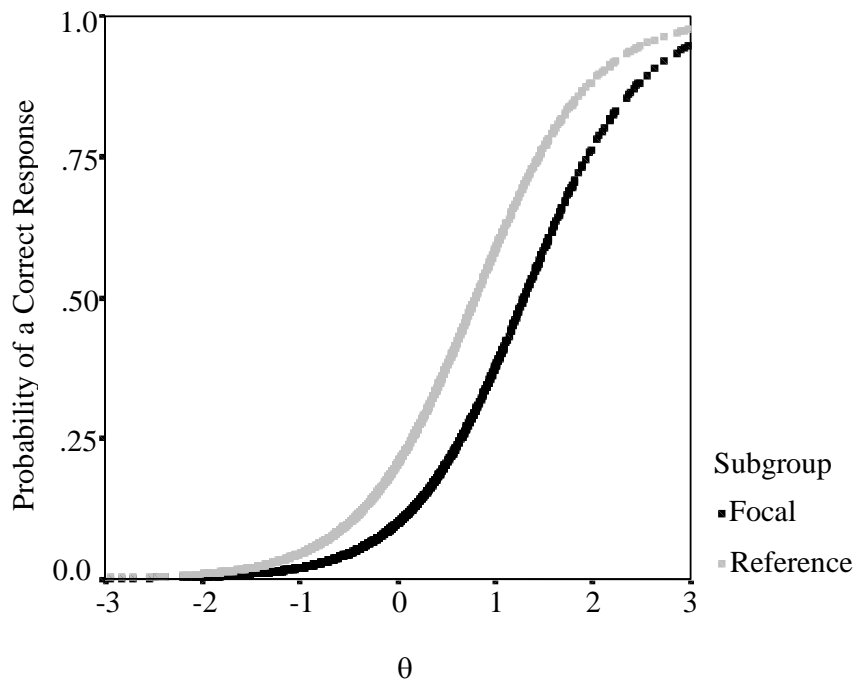


Figure 4. Uniform DIF: Reference group ( $b_i = .8, a_i=1$ ) and focal group ( $b_i = 1.3, a_i = 1$ ).

Non-Uniform DIF is a between-group difference in item discrimination (Mellenbergh, 1982). Figure 5 shows ICCs that cross. Items of this type with a lower discrimination for one subgroup relative to another reflect a difference in the psychometric quality or precision of test items (Bryant, Williamson, Wooten, & Forde, 2004). Although both types of DIF have been found in cognitive ability tests used for high-stakes decisions (Florida Department of Education, 2002; Raju, 1990; Raju, Drasgow, & Slinde, 1993), research indicates that most items showing DIF appear to be of the uniform type (Hambleton et al., 1991; Swaminathan & Rogers, 1990).

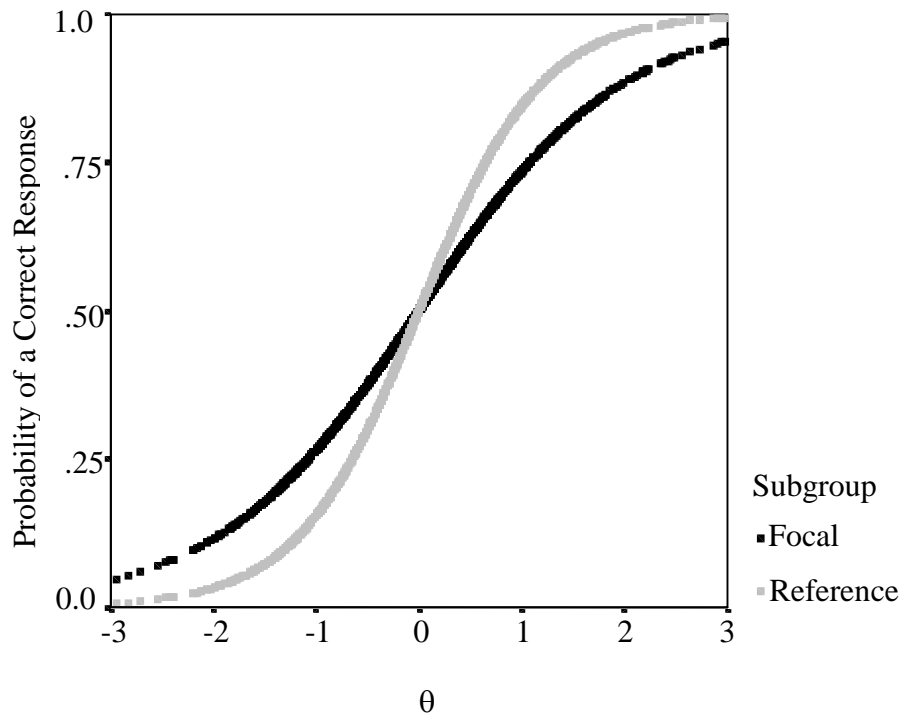


Figure 5. Non-Uniform DIF: Reference group ( $b_i = 0, a_i = 1$ ) and focal group ( $b_i = 0, a_i = .6$ ).

*Methods of detecting DIF.* A variety of methods can be used to investigate the presence of DIF in dichotomously scored items. The methods can be divided into IRT techniques (Lord, 1980; Raju, 1988, 1990; Raju et al., 1993; Roussos & Stout, 1996) and non-IRT techniques (Clauser, Mazor, & Swaminathan, 1996; Holland & Thayer, 1988; Mazor, Kanjee, & Clauser, 1995; Rudner et al., 1980; Swaminathan & Rogers, 1990). DIF has been detected using both IRT and non-IRT techniques. Some methods include logistic regression (Clauser et al., 1996; Swaminathan & Rogers, 1990), the Mantel-Haenszel statistic (Holland & Thayer, 1988), Lord's Chi-Square (Lord, 1980), SIBTEST (Shealy & Stout, 1996), and the Signed and Unsigned Area statistics (Holland & Wainer, 1993; Raju, 1988, 1990).

Some of the most widely used methods in IRT are Raju's (1988) area formulas, which provide effect size measures that are often used in simulation studies of DIF (Swaminathan & Rogers, 1990) and non-simulated DIF research (Raju, 1990). One of the most widely used non-IRT methods is the Mantel-Haenszel statistic (Holland & Thayer, 1988). Other multidimensional DIF methods include SIBTEST and logistic regression; both methods greatly improve the matching ability variables as compared to the Mantel-Haenszel method (Mazor et al., 1995; Shealy & Stout, 1996).

*Evidence of DIF.* Research on multidimensional DIF suggests that fewer items exhibit DIF as compared to unidimensional methods (Clauser et al., 1996; Mazor et al., 1995; Mazor, Hambleton, & Clauser, 1998). In simulation studies, matching on multiple subtest scores has been shown to result in fewer items exhibiting DIF (Mazor et al., 1998). Research by Mazor et al. (1995) has demonstrated that when an irrelevant dimension, e.g., verbal ability, is controlled using logistic regression, fewer items show DIF in some achievement tests. Other evidence also

suggests that simultaneously matching on several relevant, subtest scores or item-bundles results in fewer items identified as exhibiting DIF as compared to the unidimensional Mantel-Haenszel approach (Clauser et al., 1996; Shealy & Stout, 1993). This evidence supports the position that multidimensional DIF detection techniques are more able than unidimensional techniques to distinguish DIF that is due to relevant ability differences on secondary dimensions and DIF caused by construct-irrelevant variance.

Multidimensional measurement models have provided added insight into what test items measure in different subgroups (Ackerman, 1992, 1994a). Moreover, it has been posited, and evidence has supported the view, that some items show DIF because of a relevant between-group difference on a secondary trait not fully accounted for in the unidimensional composite (Shealy & Stout, 1993). When multidimensional DIF techniques are used in conjunction with multidimensional measurement models, the number of items that show DIF is somewhat reduced (Mazor et al., 1995). However, DIF has been found to occur at a rate of approximately 20% to 30% when relevant, secondary abilities are taken into account (Clauser et al., 1996). The aforementioned evidence suggests that when multidimensional measurement models are used in conjunction with multidimensional DIF detection techniques, the number of items detected as showing DIF is reduced, but construct-irrelevant factors may still be present in cognitive ability tests, which may compromise the overall validity judgment of tests used to make high-stakes decisions.

## *Predictive Bias*

*Background.* Another form of bias that has been mentioned in the *Standards* and the *Principles* concerns the prediction of performance on a criterion of interest for different subgroups (AERA, APA, & NCME, 1999; SIOP, 2003). This form of bias has emerged from earlier conceptions of single-group validity (Boehm, 1972) and differential validity. Both concepts are based on testing differences between correlation coefficients. Single-group validity requires testing a set of hypotheses for two subgroups (Subgroup 1 and Subgroup 2) for which validity coefficients are available. The first hypothesis is that the validity coefficient for Subgroup 1 is equal to zero (i.e.,  $\rho_1 = 0$ ). The second hypothesis is that the validity coefficient for Subgroup 2 is not significantly different from the validity coefficient of Subgroup 1 (i.e.,  $\rho_1 = \rho_2$ ). In addition to the first and second hypotheses, the third hypothesis is that the validity coefficient for Subgroup 2 is significantly different from zero (Boehm, 1972).

The above notion of single-group validity has been found to be untenable in the population (Bartlett, Bobko, & Pine, 1977; Linn, 1978). For example, careful evaluation of the set of hypotheses above reveals that all three hypotheses cannot simultaneously exist in the population. Therefore, single-group validity has been reduced to a sample-specific phenomenon. As an alternative to single-group validity, Humphreys (1973) proposed the test of the equality of validity coefficients as a more tenable psychological assumption. Moreover, the test of the equality of validity coefficients seems to be more consistent with the intent of regulations proposed at that time by the Equal Employment Opportunity Commission (1970). Accordingly, the subsequent version of the *Standards* (a) rejected the view of single-group validity, (b) rejected the means of detecting it, and (c) supported the position of testing the difference

between subgroup correlation coefficients (APA, 1974). Testing the difference between correlation coefficients has been subsequently rejected in favor of testing the differences in subgroup intercepts and slopes, which is now recognized as predictive bias studies (AERA, APA, & NCME, 1985, 1999). In contrast to the investigation of item bias, predictive bias studies are based on an external criterion, whereas investigations of item bias are based on an internal criterion (Camilli, 1993; Camilli & Shepard, 1994). Now details of predictive bias are discussed.

First, a definition of predictive bias is presented. Then, the method and assumptions of the model most useful in detecting predictive bias are outlined. Finally, evidence of predictive bias is briefly discussed.

*Definition of predictive bias.* The definition of predictive bias is elucidated by Cleary (1968):

A test is biased for members of a subgroup of a population if in the prediction of a criterion for which the test was designed, consistent non-zero errors of prediction are made for members of the subgroups. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of “unfair,” particularly if the use of a test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance. (p. 115)

This form of bias is assessed once data on a relevant criterion are obtained. Moreover, it is assumed that the criterion is not susceptible to measurement bias (Cascio, 1998). Despite the presence of research that challenges the regression model on social grounds, other models of bias have been proposed and investigated, e.g., the Constant Ratio model (Thorndike, 1971) and the Conditional Probability model (Cole, 1973). However, for the purpose of the present study, the focus is on the model of fairness implied by the *Standards* in evaluating differences in subgroup intercepts and slopes (Cleary, 1968; Cleary et al., 1975). This fairness model by Cleary (1968) investigates bias by means of linear multiple regression, which is described next.

*Regression model.* Consider a criterion variable represented as a vector,  $\mathbf{y}$ , and two predictor variables represented by vectors,  $\mathbf{x}_1$  and  $\mathbf{x}_2$ ;  $\mathbf{y}$  is assumed to be continuous, while  $\mathbf{x}_1$  and  $\mathbf{x}_2$  can be continuous or discrete. For an observation,  $i$ , a linear model that predicts a score  $y_i$  given scores  $x_{1i}$ ,  $x_{2i}$ , and their interaction – i.e.,  $x_{3i}$ , which is the product of mean-centered scores of  $x_{1i}$  and  $x_{2i}$  – is expressed in the following form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \quad (1)$$

or in matrix form for all of  $N$  observations,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

$\mathbf{y}$  is a vector ( $N \times 1$ ) of observations on the criterion,  $\mathbf{X}$  is a matrix ( $N \times 4$ ) of scores (the first column is a vector of ones, the second column is a vector of centered scores on  $\mathbf{x}_1$ , the third column is a vector of centered scores on  $\mathbf{x}_2$ , and the fourth column is a vector representing  $\mathbf{x}_3$ , which is the product of mean centered scores of  $\mathbf{x}_1$  and  $\mathbf{x}_2$ ),  $\boldsymbol{\beta}$  is a vector ( $4 \times 1$ ) of population

regression coefficients, i.e.,  $[\beta_0, \beta_1, \beta_2, \beta_3]'$  for the constant vector of ones and each of the three predictor variables (i.e.,  $x_1, x_2$  and the interaction,  $x_3$ ), and  $\boldsymbol{\varepsilon}$  is a vector ( $N \times 1$ ) of error terms for each of  $N$  observations.

Estimates of population regression coefficients are computed by means of ordinary least squares (Rencher, 2000) to yield

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \quad (3)$$

where  $\mathbf{b} = [b_0, b_1, b_2, b_3]'$  is a vector of estimated population regression coefficients or unstandardized beta weights (Cohen & Cohen, 1983; Darlington, 1990; Rencher, 2000). The expectation of  $\mathbf{b}$  is  $\boldsymbol{\beta}$  (Rencher, 2000). The predicted value of the criterion  $\mathbf{y}$  is given by the following equation:

$$\hat{\mathbf{y}} = \mathbf{y} - \boldsymbol{\varepsilon}. \quad (4)$$

Because of the law of expectation and Equation 4, Equation 2 can be re-written as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}, \quad (5)$$

or for an individual observation,

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i}, \quad (6)$$

where  $\hat{y}_i$  is the predicted value of the criterion for an observation,  $i$ .  $b_0$  is the estimate of the population intercept, which is the point where the predicted regression line crosses the  $y$ -axis.  $b_1$  is the population estimate of the slope of  $x_1$  or the increase in  $y$  per unit increase in  $x_1$ .  $b_2$  is the

population estimate of the slope of  $x_2$ , and  $b_3$  is the population estimate of the rate of change for the interaction or the moderating relation between  $x_1$  and  $x_2$  (Aiken & West, 1991; Stone, 1988).

There are several sets of assumptions in regression theory that involve the specification of the regression model, measurement error, and the error in prediction (Aiken & West, 1991; Darlington, 1990; Lewis-Beck, 1980). With respect to model specification, it is assumed that (a) the relation between the predictor and the criterion is linear, (b) no relevant variables have been excluded, (c) no irrelevant predictor variables have been included, and (d) the predictor and criterion have been measured reliably. The last set of assumptions dealing with the error term maintains the following: (a) the expected value of error is zero, (b) the variance of the error is constant across all levels of the predictor, i.e., homoscedasticity, (c) error terms are uncorrelated, (d) the predictors are uncorrelated with the error term, and (e) the error term is normally distributed. To the extent that the aforementioned assumptions are met, the estimators of the population parameters will be the best linear unbiased estimates (Rencher, 2000).

The multiple regression technique has been advanced as the most appropriate and recommended procedure for detecting predictive bias (AERA, APA, NCME, 1999; APA, 1980; SIOP, 2003; Stone, 1988; Stone & Hollenbeck, 1989). As applied to the regression model described above, the model to detect predictive bias can be operationalized as follows:

$$\hat{y}_i = b_0 + b_1X_{Ti} + b_2g_i + b_3(X_{Ti} \cdot g_i), \quad (7)$$

where  $\hat{y}_i$  is the predicted value of the criterion variable for observation  $i$ ;  $b_0$  is the sample estimate of the population intercept;  $b_1$  is the estimated slope coefficient for the score,  $X_T$ , which is the observed total test score;  $b_2$  is the estimated slope for subgroup membership,  $g$ , which is an

effects coded variable for subgroup membership (Aiken, West, & Krull, 1991); and  $b_3$  is the estimated slope coefficient for the interaction of test scores and subgroup membership ( $X_T:g$ ). Several situations can occur in predicting scores on a criterion: (a) a similar regression line for both subgroups, (b) uncommon regression lines due to subgroup intercept differences, and (c) uncommon regression lines due to a subgroup slope difference (Cleary et al., 1975). When using a common regression line to predict criterion scores, predictive bias occurs when (b) or (c) exists (Cascio, 1998). That is, predictive bias is present when there is either a significant difference in subgroup intercepts or subgroup slopes (Cleary et al., 1975; Schmidt & Hunter, 1974). When there is a difference in intercepts or slopes, predicting criterion scores on the basis of the majority group's regression equation may lead to the inference of unfairness because criterion scores for members of the minority group will be over-predicted (Jensen, 1980; Hartigan & Wigdor, 1989; Schmidt & Hunter, 1974). If there is a significant difference between subgroup intercepts, assuming no difference in subgroup slopes, it is evidenced in Equation 7 by a statistically significant  $b_2$  term, which means that the null hypothesis  $\beta_2 = 0$  was rejected. There would be two regression equations for each subgroup. For example, under the condition that  $b_2$  is significant,  $b_3$  is not statistically significant, and Subgroups 1 and 2 are coded as 1 and  $-1$ , respectively, then there would be different equations for the two subgroups. It would be inappropriate to use a common regression equation to predict scores on the criterion. So, different prediction equations are needed. The prediction equation for Subgroup 1 is

$$\hat{y}_i = b_0 + b_2 + b_1 X_{Ti}, \quad (8)$$

and the prediction equation for Subgroup 2 is

$$\hat{y}_i = b_0 - b_2 + b_1 X_{Ti} \quad (9)$$

As an example, Subgroup 1 is the reference group (Anglo-American) to which all other subgroups are compared, and Subgroup 2 is the focal group (African-American) of interest. Equations 8 and 9 are general representations of reference and focal group equations. An illustration is presented in Figure 6. This example of differential prediction shows that subgroup intercepts are different, but subgroup slopes are equal. In other words, the means on the test are different but the relation between the test (i.e., the predictor) and the criterion is the same (Cascio, 1998).

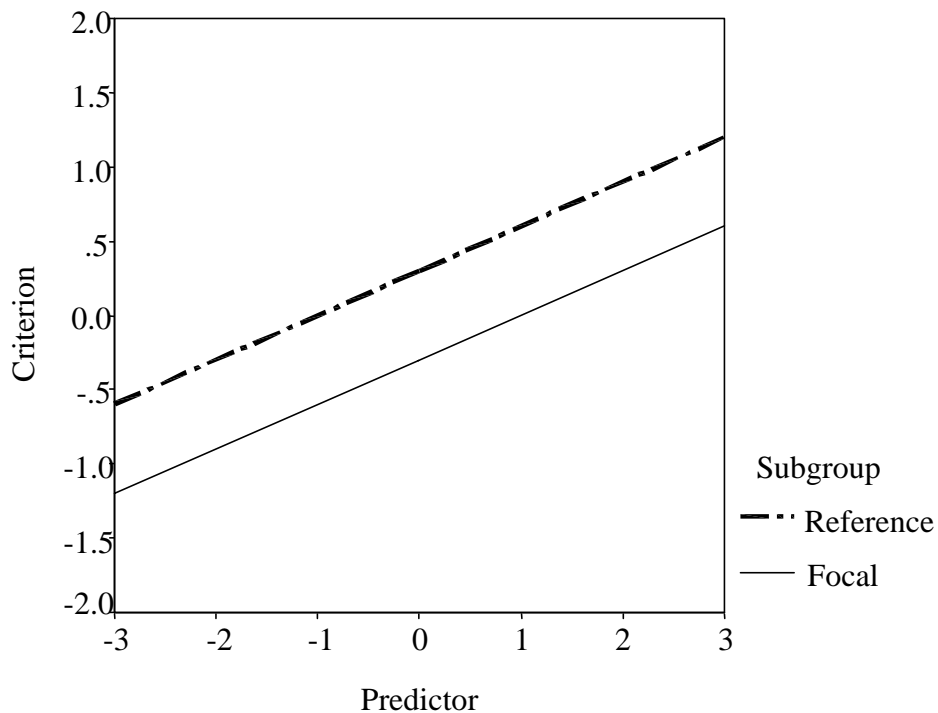


Figure 6. Predictive bias with different subgroup intercepts and equal slopes.

When there is a difference in subgroup slopes, the strength of the relation between the predictor and the criterion is not the same for both subgroups (Cleary et al., 1975). This is indexed by a statistically significant  $b_3$  term, irrespective of the statistical significance of  $b_2$ . Under this condition, the estimate of the slope parameter for the test score ( $X_T$ ) and subgroup interaction will be different, thus leading to two regression equations quite different from Equations 8 and 9. Equation 7 is also used in another example. Assume that  $b_3$  is statistically significant,  $b_2$  is not significant, and Subgroups 1 and 2 are coded as in the previous example. The regression equation for Subgroup 1 is

$$\hat{y}_i = b_0 + (b_1 + b_3) X_{Ti}, \quad (10)$$

while the regression equation for Subgroup 2 is

$$\hat{y}_i = b_0 + (b_1 - b_3) X_{Ti}. \quad (11)$$

An illustration of the difference in subgroup slopes is given in Figure 7. As can be seen from the graph, the slope for the focal group (Equation 11) is relatively flat compared to that of the reference group (Equation 10). Subgroup intercepts are not relevant (Cascio, 1998).

When there are differences of the same magnitude and in the same direction on the predictor and the criterion, it is expected that a situation will occur where there is a common regression line that predicts equally for both reference and focal groups. This is illustrated in Figure 8. The graph shows that both subgroups have the same intercept and slope; therefore, separate regression lines are not needed.

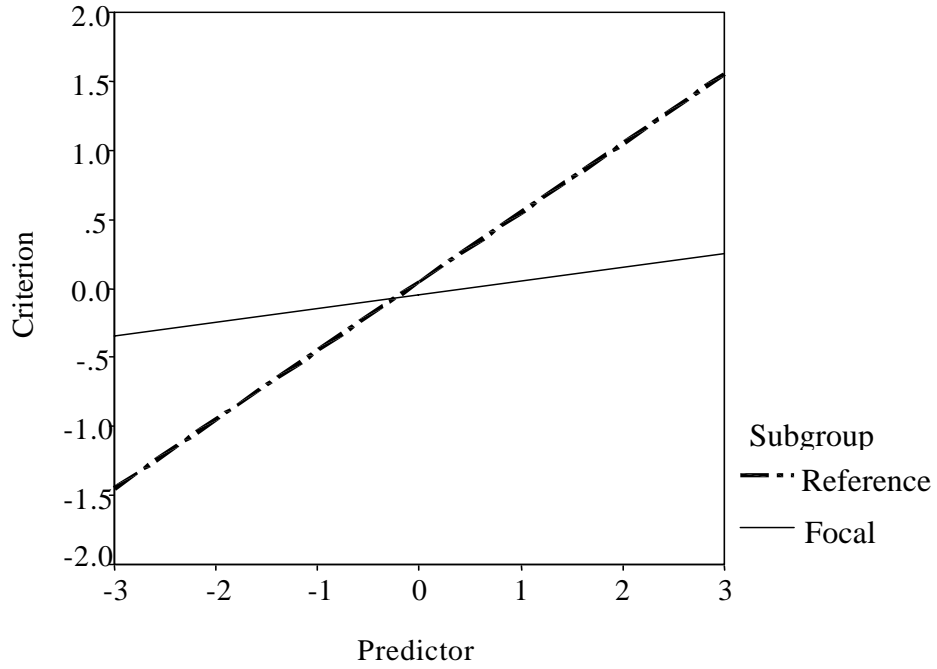


Figure 7. Predictive bias with different subgroup slopes.

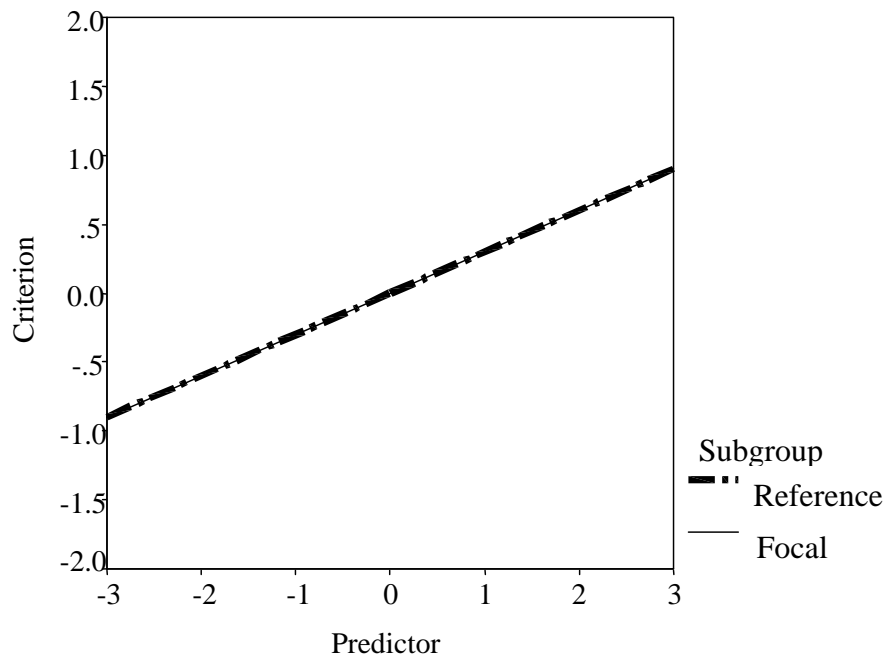


Figure 8. Equality of subgroup intercepts and slopes.

*Evidence of predictive bias.* In terms of predictive bias, evidence suggests that there are differences in subgroup intercepts with little evidence of slope differences (Casio, 1998; Hartigan & Wigdor, 1989; Hunter & Schmidt, 2000; Jensen, 1980; Schmidt & Hunter, 1974). Two committees have been commissioned by the National Academy of Science to investigate predictive bias in cognitive ability tests used for federal government hiring (Hartigan & Wigdor, 1989). In a report on the analysis of 72 studies with at least 50 African-Americans and 50 non-minorities, two studies have shown the presence of slope differences (i.e., approximately 3%). Twenty-six of the remaining studies (i.e., approximately 34%) have shown evidence of intercept differences (Hartigan & Wigdor, 1989). A review of over 1,000 predictive bias studies by Bartlett, Bobko, Mosier, and Hannan (1978) indicates that in employment testing used for selection, some tests show a significant difference in subgroup intercepts approximately 24% of the time; however, the bias is interpreted to be against the majority or reference group.

The empirical research over the past couple of decades suggests that cognitive ability tests overpredict the performance of minority group members (Bartlett et al., 1978; Hunter et al., 1984; Jensen, 1980). Evaluation of these studies suggests that there is little evidence of tests being biased, in the predictive sense, against minorities. In the vast majority of conditions, differential prediction has been found when there is a difference in subgroup means in favor of the majority group (Schmidt & Hunter, 1974). There have been concerns, however, about the substantial false rejection rate of minority members in employment decisions due to the fallible testing technology (Hartigan & Wigdor, 1989), which can be considered as one of the adverse consequences of decisions based on test score interpretations (AERA, APA, & NCME, 1999).

### *Link Between DIF and Predictive Bias*

Although there have been very few, if any, empirical investigations of both item bias and predictive bias, researchers have speculated that if item bias exists against a subgroup (e.g., African-Americans), then predictive bias will occur against the same subgroup (Hunter & Schmidt, 2000; Hunter et al., 1984). Hunter et al. (1984) state:

If only certain items were biased, then the test as a whole would still be as valid for blacks as for whites considered separately. However, the test scores of blacks would be systematically lower than those of whites of the same ability level because blacks would miss the biased items. If it were true that tests underestimate black ability, then it would follow that test scores would underpredict black performance on the job. This, in turn, leads to the prediction that if tests were biased against blacks, then the regression line for blacks would lie above the regression line for whites. The data show just the reverse to be true. (pp. 49-50)

In this statement, there is an issue of a test still being *valid* if items are biased. Hunter et al. (1984) imply that validity is determined primarily by the relation test scores have with an external criterion. A couple of facts about these two types of bias deserve mentioning. First, item and predictive bias investigations answer different questions: (a) item bias studies, as mentioned above, are concerned with questions related to the internal structure of the test, and (b) predictive bias studies answer questions about the relation that test scores have with an external criterion. Notwithstanding suppositions of Hunter and others (Hunter & Schmidt, 2000; Hunter et al.,

1984), the relations among DIF, total test scores, and the prediction of a criterion have not been fully delineated (Jones & Applebaum, 1989). So, conclusions about the relation between item bias and predictive bias are speculative at best. However, if a relation were to be specified, psychometric factors that influence the detection of predictive bias may give insight into the relation between measurement bias at the item level (i.e., DIF) and predictive bias.

Although speculation about the relation between item bias and test bias has persisted over decades (Hunter & Schmidt, 2000; Hunter et al., 1984), no empirical evidence has shown how item bias against a subgroup would lead to predictive bias against it. An attempt is made in the present study to clarify conditions that would cause predictive bias, which may lead to testable hypotheses. To facilitate interpretation, assume that a regression model is used that excludes the term for the interaction of group membership and test scores. Equation 7 without the interaction is now written in standardized form as follows:

$$z_{\hat{y}_i} = B_1 z_{X_{Ti}} + B_2 z_{g_i}, \quad (12)$$

where  $z_{\hat{y}_i}$ ,  $z_{X_{Ti}}$ , and  $z_{g_i}$  are standardized values of the criterion, predictor (i.e., the test), and subgroup membership for person  $i$ ,  $B_1$  is the estimated standardized regression coefficient for the predictor, and  $B_2$  is the estimated standardized regression coefficient for subgroup membership. In this example, predictive bias is indicated by a statistically significant standardized regression coefficient,  $B_2$ , which can be computed using correlations between the criterion and subgroup membership ( $r_{yg}$ ), the predictor and subgroup membership ( $r_{gx}$ ), and the predictor and criterion ( $r_{xy}$ ):

$$B_2 = [r_{yg} - (r_{yx})(r_{gx})] / [1 - (r_{gx})^2]. \quad (13)$$

Significance testing of the null hypothesis for the population coefficient ( $\beta_2$ ) is given as

$B_2/SE(B_2)$ , where  $SE(B_2)$  is the standard error of  $B_2$  given by

$$SE(B_2) = \{[1 - (B_1 r_{yx} + B_2 r_{yg})] / [(1 - r_{gx}^2)(N - k - 1)]\}^{1/2}, \quad (14)$$

where  $N$  is the number of observations, and  $k$  is the number of predictors (in this case,  $k = 2$ ).

Equations 13 and 14 are used in this study to illustrate several scenarios and derive hypothesis about the relation between predictive bias and DIF.

*Scenario 1.* In Equation 13, if  $r_{yg} = r_{gx} = 0$  and  $r_{yx} = .3$ , then  $B_2 = 0$ , thus there is no predictive bias. In other words, if subgroup membership is unrelated to the criterion and subgroup membership is unrelated to the test, then it follows that there is no predictive bias. An illustration of Scenario 1 appears in Figure 8.

*Scenario 2.* Now assume that the reference group is coded 1 and the focal group is coded -1. When subgroup membership is correlated with the test (or a criterion), the correlation coefficient between subgroup membership and the predictor (or a criterion) is positive if the reference group has a higher mean score on the predictor (or a criterion) than the focal group. The correlation is negative if the focal group has a higher mean score than the reference group. If  $r_{gx} = 0$ , then  $B_2 = r_{yg}$ , irrespective of the value of the validity coefficient ( $r_{yx}$ ) in Equation 13. In other words, when there is no association between subgroup membership and predictor scores, then the direction, magnitude, and significance of  $B_2$  are determined primarily by the association between subgroup membership and the criterion,  $r_{yg}$ . Thus, a mean subgroup difference on the criterion has major implications for predictive bias, especially when there is no difference on the predictor. For instance, if  $r_{yx} = .3$ ,  $r_{gx} = 0$ , and  $r_{yg} = .3$ , predictive bias may result in this situation

with  $B_2 = .3$ , assuming a sample of approximately 50 or more. Regression lines indicative of this scenario are shown in Figure 6.

When  $r_{gx} = 0$ , the value of  $SE(B_2)$  is lower as compared to the situation where the absolute value of  $r_{gx}$  is greater than zero, holding all else constant. Moreover,  $SE(B_2)$  decreases as the sample size, represented in the denominator of Equation 13, increases. As  $SE(B_2)$  decreases, the chance of rejecting the null hypothesis (i.e.,  $\beta_2 = 0$ ) increases. Thus, when the relation between subgroup membership and the criterion is positive, the intercept of the reference group (1) is higher than the intercept for the focal group (-1) in both the standardized and unstandardized equations. If the slopes are the same, then this represents a situation often found in the literature where the minority group regression line is lower than the majority group regression line (Hunter & Schmidt, 2000; Hunter et al., 1984; Jensen, 1980; Schmidt & Hunter, 1974).

*Scenario 3.* The coding scheme is the same as in Scenario 2. If it is assumed in Equation 13 that the validity coefficient ( $r_{yx}$ ) is positive and the reference group has a higher mean score on the criterion than the focal group (i.e.,  $r_{yg}$  is positive) due to a true ability difference, systematic error, or a combination of both, different outcomes may occur. Two possible outcomes are considered. First, it is possible that the focal group can have a lower mean score on the test than the reference group ( $r_{gx}$  is positive) and the regression model will still produce a line with a higher intercept for the reference group as compared to the focal group. In other words, it is theoretically plausible that the focal group can have a mean score on the test that is lower than the mean score of the reference group (due to a true difference, measurement bias, or a

combination of both) and an intercept difference or predictive bias will occur with a lower intercept for the focal group as compared to the reference group, holding all else constant. A graphical representation for this situation is also similar to the illustration for predictive bias (Figure 6), which is commonly interpreted as predictive bias against the majority group (Chung-Yan & Cronshaw, 2002; Drasgow, 1982, 1984; Kraiger, Ford, & Schechtman, 1986).

Moreover, it is possible under the same set of circumstances described in Scenario 3 that predictive bias will not occur for either subgroup. For this second example, assume the following: (a) the correlation between subgroup membership and scores on the criterion ( $r_{yg}$ ) is .179, (b) the correlation between scores on the test and scores on the criterion measure ( $r_{yx}$ ) is .397, and (c) the correlation between subgroup membership and scores on the test ( $r_{gx}$ ) is .450. The standardized regression coefficient for the test ( $B_1$ ) is estimated to be .397. Although the difference between subgroups is larger on the test as compared to the difference on the criterion, as indicated by the larger point-biserial correlation coefficient (.179 versus .450), the standardized regression coefficient for subgroup membership ( $B_2$ ) is estimated to be approximately 0. In essence, both subgroups would be adequately represented by one regression line (Figure 8). In both examples within this scenario, the test is not biased against the focal group in the predictive sense, although true ability differences on the test and criterion are possibly obfuscated with systematic errors occurring against the focal group.

### *Hypotheses*

Because a necessary condition for predictive bias is a subgroup difference in mean scores on either the predictor or the criterion, manipulating factors that influence mean differences on a

predictor (i.e., a test) and a criterion may give insight into how DIF may influence the prediction of a criterion. Based on the above review on DIF, two DIF-related factors may influence mean subgroup differences beyond true ability differences: (a) differences in item response functions, and (b) the percentage of items that exhibit DIF. From Equations 13 and 14, several factors may influence the detection of predictive bias: (a) a true mean subgroup difference on the predictor, (b) a true mean subgroup difference on the criterion, (c) the sample size of subgroups, and (d) the validity coefficients. No research has attempted to establish an empirical link between item bias and predictive bias. Several of these factors were investigated in the current study. Thus, the following hypotheses were tested:

*Hypothesis 1.* As the percentage of DIF against one subgroup increases, the detection of predictive bias increases.

DIF increases the probability of getting an item right for members of the reference group and decreases the probability of getting an item right for members of the focal group with the same ability, which in a cumulative sense may increase the difference in subgroup means beyond that due to true ability differences. This represents a condition that is similar to the first example described in Scenario 3. So, as the percentage of DIF increases the detection of predictive bias should increase. If true ability differences are obfuscated by differences due to DIF, it is possible that predictive bias due to mean score differences may increase the extent of predictive bias. This hypothesis is evaluated both by the detection of predictive bias and the direction of predictive bias (i.e., against the focal group or against the reference group).

*Hypothesis 2.* As the effect size of DIF against one subgroup increases, the detection of predictive bias increases.

Because uniform DIF (Figure 4) can artificially and consistently shift probabilities of getting the item correct for some members of the reference group relative to members of equal ability in the focal group, it is reasonable to suspect that DIF of this type, which is often found in DIF analyses, may influence the difference in subgroup means on a test in addition to any other true differences that might exist (see the first example in Scenario 3). Specifically, it is plausible that the size of DIF at the item level can increase the difference in mean scores between subgroups by creating item-level disparities in the distance between subgroup points of maximum item information. The point of maximum information for each subgroup is the item difficulty ( $b_i$ ). In this circumstance, as the size of uniform DIF increases against one subgroup, the mean difference between subgroups should artificially increase, holding all else constant. For small amounts of DIF, the increase in the difference in subgroup means will not be as great as large amounts of DIF. When a mean subgroup difference on the criterion is not the same in magnitude as the mean difference on the predictor, predictive bias may result as in Scenario 3. This hypothesis is evaluated both by the detection of predictive bias and the direction of predictive bias (i.e., against the focal group or against the reference group).

*Hypothesis 3.* As subgroup sample size increases, the detection of predictive bias increases.

Evidence from both simulated and non-simulated studies of multiple regression indicates that as sample size increases, as described in Scenario 2, the rate of rejecting the null hypothesis increases due to the fact that the standard error of the estimated regression coefficient decreases as the number of observations increases (Darlington, 1990; Stone-Romero & Anderson, 1994). In the current study, as the sample size of subgroups increases, the standard error of the

regression coefficient for the group variable or the group by ability interaction should decrease, thus increasing the chance of rejecting the null hypothesis.

*Hypothesis 4A.* As the mean difference on the predictor increases, predictive bias increases.

*Hypothesis 4B.* As the mean difference on the criterion increases, predictive bias increases.

As described in Scenario 2, when there are true differences on the predictor or the criterion, the chance of detecting predictive bias should increase as the ability difference between the subgroups increases. Because there is some evidence (Chung-Yan & Cronshaw, 2002; Ford, Kraiger, & Schechtman, 1986) to suggest that there are differences in ability on some objective measures of criteria and predictors (Hough, Oswald, & Ployhart, 2001), this variable is important to manipulate as a means of investigating the relation between measurement bias at the item level and predictive bias at the test score level.

## METHOD

The method used to evaluate Hypotheses 1-4 was a simulation study. The Python (Lutz & Ascher, 1999) programming language was used to simulate all data. The independent variables were those hypothesized to influence (a) mean subgroup differences and (b) the detection of predictive of bias. Two of the variables were factors that influence the internal structure of the test and were hypothesized to influence subgroup means on the predictor: (a) percentage of DIF and (b) effect size of DIF. Four of the independent variables were hypothesized to influence the detection of predictive bias: (a) sample size, (b) validity coefficient or the correlation between the test (predictor) and the criterion within each subgroup, (c) true mean subgroup differences on the predictor, and (d) true mean subgroup differences on the criterion. After delineating fixed conditions, a description of the test design is given. Next, the independent and dependent variables are described, which is then followed by a description of the data analytic procedures.

### *Fixed Conditions*

*Test length.* The number of items used in the simulated test was held constant at 60 items. The rationale was due to the fact that for most tests given in employment and educational contexts, a large number of items are needed in order to improve the psychometric quality of the test; this is consistent with the means by which test reliability is increased in CTT (Nunnally & Bernstein, 1994).

*Number of replications.* Past research showed that the number of replications for stable parameter estimation is rather large (Stone & Hollenbeck, 1989; Stone-Romero & Anderson,

1994). Thus, each condition was replicated 500 times in the 3 x 2 x 4 x 2 x 3 x 3 design described below.

*Direction of the mean subgroup difference.* For the purpose of this study, when a mean subgroup difference occurred, it was in favor of the reference group. This was done to reflect what has been reported most often in the employment and educational literature with respect to mean subgroup differences.

### *Test Design*

Because of the multidimensional nature of cognitive ability tests used for employment selection (Hunter & Schmidt, 2000; SIOP 2003), the multidimensional two-parameter logistic model, M2PL, (Reckase, 1985; Reckase & McKinley, 1991) was used to create item responses and to embed DIF that was independent of differences in the subgroup trait distributions. The model (Reckase, 1985) is expressed in following general form:

$$P_i(\boldsymbol{\theta}_j) = \{1 + \text{Exp}[-D(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)]\}^{-1}, \quad (15)$$

where  $D$  is equal to a scaling constant 1.7,  $\mathbf{a}_i$  is a vector of  $k$  discrimination parameters for item  $i$ ,  $[a_{1i}, a_{2i}, \dots, a_{ki}]'$ ,  $k$  is the number of dimensions,  $\boldsymbol{\theta}_j$  is a vector of  $k$  ability parameters for person  $j$ ,  $[\theta_{1j}, \theta_{2j}, \dots, \theta_{kj}]'$ , and  $d_i$  is a scalar related to difficulty.

*Item parameters.* For the purpose of this study, a two-dimensional ability test was simulated according to the model of Equation 15. Logistic models of this kind have been used to design tests such as the ACT (Ackerman, 1994b; Reckase, 1997) and the ASVAB (Segall, 1996). They also have been used to evaluate the dimensionality of multidimensional tests, such as the

Law School Admissions Test (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). The test in this study was designed so that scale score consistency (Ackerman, 1994b) was maintained across levels of the number correct score, i.e., a score which is correlated most highly with the direction of best measurement of the composite of  $\theta_1$  and  $\theta_2$ . The direction of best measurement was defined as the direction in the latent ability space where information or precision is at a maximum. The direction of best measurement for the item was obtained using the following algorithm: (a) find the length of  $\mathbf{a}_i$ , which is the square root of the sum of squared elements of  $\mathbf{a}_i$ , (b) divide each element in  $\mathbf{a}_i$  by the length of  $\mathbf{a}_i$  to get the cosine of the angle between the item vector and each of the  $\theta$  axes, and (c) compute the arc cosine of each element in  $\mathbf{a}_i$  to determine the item's direction from each of the  $\theta$  axes. For example, if an item has a discrimination vector of  $\mathbf{a}_i = [.974, .229]'$ , then the direction of best measurement for this item would be 13.4 degrees from  $\theta_1$  and 76.6 degrees from  $\theta_2$ . Thus, the item would measure more of  $\theta_1$  as compared to  $\theta_2$ . The direction of best measurement for the theta composite ( $\theta_c$ ) was the direction in the ability space that provided the most psychometric information (Ackerman, 1994b).

Although any direction of measurement for the  $\theta_c$  can be specified in a multidimensional model, it was assumed hypothetically that a job or curriculum analysis revealed that  $\theta_1$  was more important than  $\theta_2$ , and both abilities were putatively related to some job (or educational) criterion. In addition, it was assumed that the importance weights of the two abilities were determined by a group of subject matter experts. Thus, the measurement direction for the composite score was specified as  $\mathbf{u} = [.8366, .5477]'$  or 33 degrees from  $\theta_1$  and 57 degrees from  $\theta_2$ . This direction was represented in a Cartesian graph as a line that passes through the origin (0,

0) and the point  $\mathbf{u} = [.8366, .5477]$  in a two-dimensional ability space. The validity sector of measurement (Ackerman, 1994a) around the specified direction was set at 20 degrees in either direction; items were selected that fit within the validity sector. The parameters for the 60-item test are listed in Table 3. See Figure 9 for a geometric representation of the items. Each point in Figure 9 represents the location on  $\theta_1$  and  $\theta_2$  where the item has its greatest discriminating potential. These values were computed using  $\theta_{\max}$  in Table 2 for the M2PL model. The  $\theta_c$  was computed as  $\mathbf{u}'\boldsymbol{\theta}$  or  $\theta_c = .8366(\theta_1) + .5477(\theta_2)$ . Ackerman (1994b) provided formulas for computing information and standard errors of estimation for  $\theta_c$ . See Figures 10 and 11 for the test information function and the standard error of the composite.

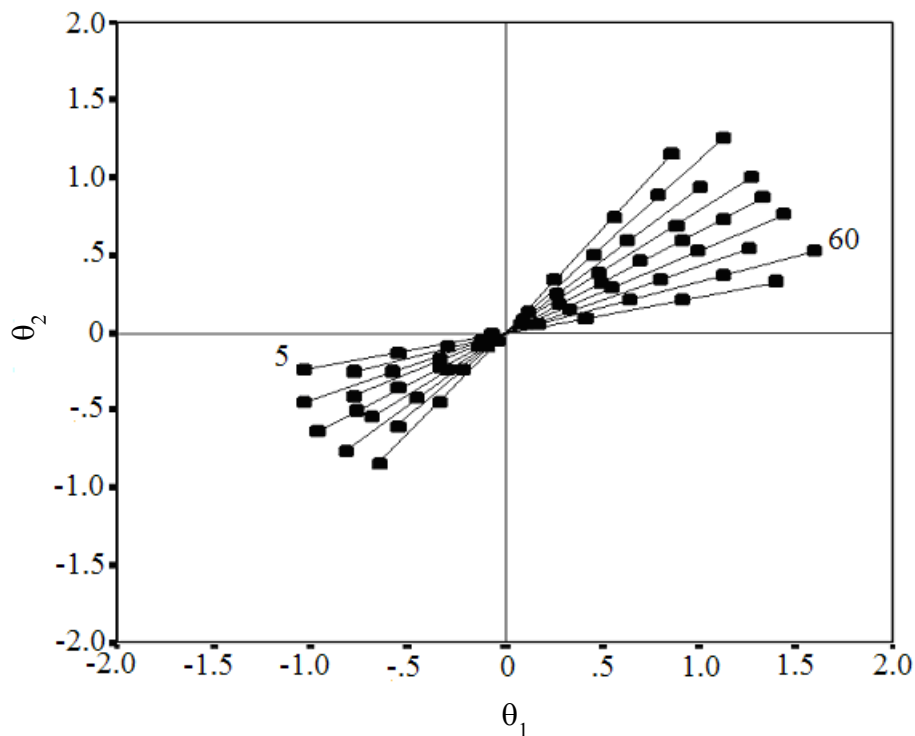


Figure 9. Geometric representation of items in two dimensions.

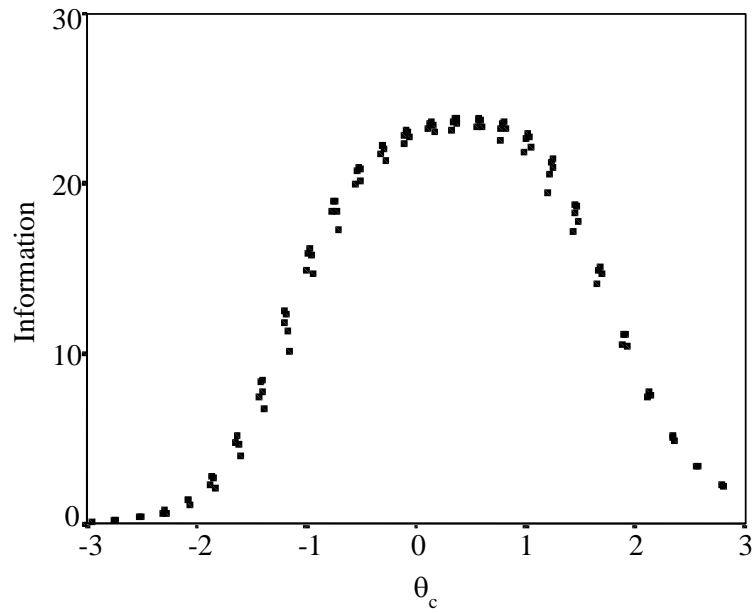


Figure 10. Test information for the theta composite,  $\theta_c$ .

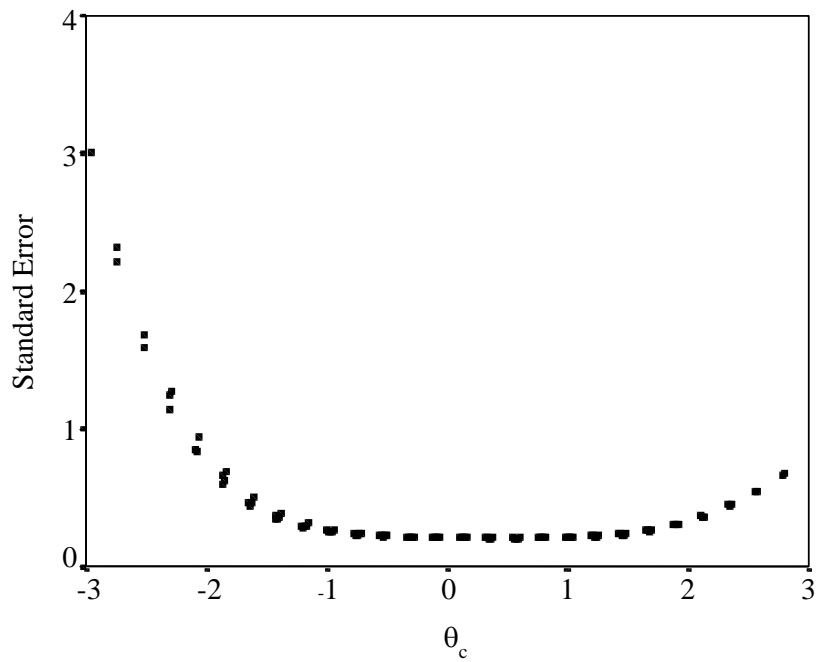


Figure 11. Standard error of estimation for the theta composite,  $\theta_c$ .

*Item responses.* The parameters of the 60 items and the ability distributions of reference and focal groups, described below, were used in the M2PL model to produce item probabilities in the range of zero to one. In order to simulate a dichotomous response to an item, each probability computed with the generated abilities  $[\theta_1, \theta_2]$  was compared to a randomly generated number from a uniform distribution (0, 1). If the randomly generated number was less than or equal to the probability, then the item was scored correct (1). If the randomly generated number was greater than the probability, the item was scored as incorrect (0). After the 60 item responses for each person within each simulation were generated, the 60 items were summed to produce a total score,  $X_T$ . This total score was the predictor in the multiple regression model used to detect predictive bias in each simulation (see Equation 7). Although subgroup proportions had been found to influence the rate of detecting subgroup slope differences (Stone-Romero, Alliger, & Aguinis, 1994), the proportion of participants in each of the two subgroups was equal, and the number in each subgroup was used to create the effects coded variable ( $g$ ) in the regression model. The algorithm used for simulating the data is provided in Appendix B.

### *Independent Variables*

There were six manipulations in this study: (a) percentage of DIF (0%, 15%, and 30%), (b) size of DIF (.3, .6, and .9), (c) subgroup sample size (35, 70, and 105 per group), (d) true validity coefficient ( $\rho_{xy} = .3$  and  $.5$ ), (e) true mean subgroup difference on the predictor (0, .33, .66, and 1 standard deviation,  $SD$ ), and (f) true mean subgroup differences on the criterion (0 and  $.35 SD$ ). Each condition in the  $3 \times 3 \times 3 \times 2 \times 4 \times 2$  design was simulated 500 times.

Table 3.

Item Parameters for the Multidimensional 2-Parameter Logistic Model

Item	$a_{i1}$	$a_{i2}$	$d_i$	$MDIFF_i$	Item	$a_{i1}$	$a_{i2}$	$d_i$	$MDIFF_i$
1	1.673	1.095	2.50	-1.25	16	1.673	1.095	1.00	-0.50
2	1.093	1.027	1.88	-1.25	17	1.179	0.928	0.75	-0.50
3	1.379	0.591	1.88	-1.25	18	1.322	0.709	0.75	-0.50
4	0.599	0.801	1.25	-1.25	19	0.666	0.746	0.50	-0.50
5	0.974	0.229	1.25	-1.25	20	0.950	0.313	0.50	-0.50
6	1.673	1.095	2.00	-1.00	21	1.673	1.095	0.50	-0.25
7	1.179	0.928	1.50	-1.00	22	1.093	1.027	0.38	-0.25
8	1.322	0.709	1.50	-1.00	23	1.379	0.591	0.38	-0.25
9	0.666	0.746	1.00	-1.00	24	0.599	0.801	0.25	-0.25
10	0.950	0.313	1.00	-1.00	25	0.974	0.229	0.25	-0.25
11	1.673	1.095	1.50	-0.75	26	1.673	1.095	0.00	0.00
12	1.093	1.027	1.13	-0.75	27	1.179	0.928	0.00	0.00
13	1.379	0.591	1.13	-0.75	28	1.322	0.709	0.00	0.00
14	0.599	0.801	0.75	-0.75	29	0.666	0.746	0.00	0.00
15	0.974	0.229	0.75	-0.75	30	0.950	0.313	0.00	0.00

Note.  $a_{i1}$  is the discrimination parameter for  $\theta_1$ ,  $a_{i2}$  is the discrimination parameter for  $\theta_2$ ,  $MDIFF_i$  is multidimensional difficulty, and  $d_i = -MDIFF_i [(a_{i1})^2 + (a_{i2})^2]^{1/2}$ .

Item	$a_{i1}$	$a_{i2}$	$d_i$	$MDIFF_i$	Item	$a_{i1}$	$a_{i2}$	$d_i$	$MDIFF_i$
31	1.673	1.095	-0.50	0.25	46	1.673	1.095	-2.00	1.00
32	1.093	1.027	-0.38	0.25	47	1.179	0.928	-1.50	1.00
33	1.379	0.591	-0.38	0.25	48	1.322	0.709	-1.50	1.00
34	0.599	0.801	-0.25	0.25	49	0.666	0.746	-1.00	1.00
35	0.974	0.229	-0.25	0.25	50	0.950	0.313	-1.00	1.00
36	1.673	1.095	-1.00	0.50	51	1.673	1.095	-2.50	1.25
37	1.179	0.928	-0.75	0.50	52	1.093	1.027	-1.88	1.25
38	1.322	0.709	-0.75	0.50	53	1.379	0.591	-1.88	1.25
39	0.666	0.746	-0.50	0.50	54	0.599	0.801	-1.25	1.25
40	0.950	0.313	-0.50	0.50	55	0.974	0.229	-1.25	1.25
41	1.673	1.095	-1.50	0.75	56	1.673	1.095	-3.00	1.50
42	1.093	1.027	-1.13	.075	57	1.179	0.928	-2.25	1.50
43	1.379	0.591	-1.13	0.75	58	1.322	0.709	-2.25	1.50
44	0.599	0.801	-0.75	0.75	59	0.666	0.746	-1.50	1.50
45	0.974	0.229	-0.75	.075	60	0.950	0.313	-1.50	1.50

Note.  $a_{i1}$  is the discrimination parameter for  $\theta_1$ ,  $a_{i2}$  is the discrimination parameter for  $\theta_2$ ,  $MDIFF_i$  is multidimensional difficulty, and  $d_i = -MDIFF_i [(a_{i1})^2 + (a_{i2})^2]^{1/2}$ .

*Predictor difference.* The composite score, defined as the sum of the item responses created from item parameters and ability distributions, was the predictor. The difference variable is represented as the difference in the  $\theta_c$  created from the ability distributions for each subgroup. The reference group had an ability distribution with the same mean vector and variance-covariance matrix,  $\Sigma$ , over all conditions, i.e.,  $\mu = [0, 0]$  and  $\Sigma$  was an identity matrix.

There were four levels of the predictor difference variable. For the first level, there was no mean difference between reference and focal groups on the first or second dimension. So, the focal group had a mean vector and variance-covariance matrix equal to the reference group. This represented a condition where there was no overall difference in the composite score. The second level represented a composite difference of  $-.333 SD$ . Thus, the focal group had means on the first and second dimensions of  $-.333$  and  $-.098$ , respectively. When weighted by the direction of  $u = [.8366, .5477]$ , the composite yielded an approximate mean difference of  $.333$ . The third level represented a composite difference of  $.666$ ; thus the focal group had means on the first and second dimensions of  $-.666$  and  $-.199$ , respectively. The fourth level represented a composite difference of one  $SD$ , which was computed as the composite of  $-1$  and  $-.298$  on the first and second dimensions, respectively, for the focal group in the direction  $u$ . The one  $SD$  difference represented the worst possible condition of observed mean differences in test scores reported in the literature.

*Criterion difference.* There were two levels of mean subgroup differences on the criterion:  $\mu = 0$  or  $\mu = .35$ . The between-group difference in criterion means was zero (0) in the first condition. Although recent meta-analytic work in the employment literature suggests that mean differences between subgroups on criteria (e.g., supervisor job ratings) are as high as  $.35$

*SD* in favor of reference group members (e.g., Anglo-American), there is considerable evidence that these criteria are biased against focal group members (Hartigan & Wigdor, 1989), which violates the assumption of the regression model used to test for predictive bias (Cascio, 1998; Darlington, 1990; Lewis-Beck, 1980). Other meta-analytic research suggests that objective criteria showed little or no evidence of subgroup differences in criterion means (Chung-Yan & Chronshaw, 2002; Kraiger, Ford, & Schechtman, 1986).

However, for the purpose of this study, the criterion was assumed to be free of bias. To represent no difference in subgroup means, the criterion was randomly distributed with  $\mu = 0$  and  $\sigma = 1$  for both reference and focal groups in the first condition. For the second condition, the criterion was randomly distributed with a  $\mu = -.35$  and  $\sigma = 1$  for the focal group. The distribution of the criterion for the reference group was similar to the first condition.

*Validity coefficient.* The manipulation of the validity coefficient consisted of only two levels. Most validity coefficients reported in the employment literature have an upper limit of approximately .5, accounting for 25% of the variance (Hattrup & Schmitt, 1990). Thus, for the purpose of this study, the two levels of the validity coefficients were .3 and .5. It should be noted that the manipulation was such that the composite measure correlated with the criterion at the specified level. This was done by weighting the two abilities by the direction of best measurement such that the composite, once created, would relate to the construct in the degree specified according to the validity coefficient manipulation. The base condition with no mean difference on the criterion and predictor served as a manipulation check of this variable. Thus, for the  $\rho = .3$  validity condition, the correlation between  $\theta_1$  and the criterion was .2744; the correlation between  $\theta_2$  and the criterion was .1212. For the  $\rho = .5$  condition, the correlation

between  $\theta_1$  and the criterion was .4574; the correlation between  $\theta_2$  and the criterion was .2021. These relations were created by generating the criterion and both ability distributions described above by a multivariate normal distribution with the mean vector specified according to the predictor and criterion difference manipulations and variance-covariance matrix specified in accordance with the validity coefficient manipulation.

*Sample size.* Because sample size was found to influence the rate of rejecting the null hypothesis in regression analysis, the manipulation of this variable involved three levels. The proportion in each subgroup was constant (i.e., equal samples sizes were maintained). The manipulations were 35 per group, 70 per group and, 105 per group. Previous research showed that regression analysis had a varied range of power to reject the null hypothesis under the aforementioned sample sizes (Cohen, 1988).

*Percentage of DIF.* There were three levels of DIF: no DIF, low DIF, and high DIF. Because it was shown in both multidimensional and unidimensional DIF detection research that the rate of DIF in cognitive ability tests can be as high as 30% of all test items, there were three manipulations in this condition. The 0% condition was the no DIF condition and represented the best-case scenario. The DIF items were randomly selected. The low DIF condition was 15% of the items (i.e., nine items: 10, 12, 19, 33, 37, 43, 55, 57, and 58). The high DIF condition had 30% of the items showing DIF (18 items: 5, 6, 10, 12, 13, 19, 21, 24, 28, 30, 33, 37, 41, 42, 43, 55, 57, and 58). The DIF items were designed to be more difficult for the focal group holding ability constant. This was done primarily as a means of assessing the influence that DIF has on the rate of detecting predictive bias against the focal group.

*Effect size of DIF.* There were three levels of effect size in this study: small, medium, and large. In the IRT metric of the area between ICCs, Raju (1988) had specified small as .3, medium as .6, and large as .9 or greater. This metric was applied to the multidimensional case by adding a small, medium, and large effect size to the multidimensional item difficulty parameter, *MDIFF*. See A1-A4 in Appendix A for a description of the multidimensional IRT statistics.

### *Dependent Variable*

*Predictive bias.* For each data simulation in each condition, Equation 7 was used to detect predictive bias. Detection of predictive bias was treated as a dichotomous variable. For each simulation within each condition, if predictive bias was detected either by a difference in subgroup intercepts or in subgroup slopes, the observation was coded as one (1). If predictive bias was not detected, the observation was coded as zero (0). There were two predictive bias variables: One for the detection of predictive bias against the focal group and another for the detection of predictive bias against the reference group. All hypotheses were evaluated with each of the predictive bias variables.

### *Data Analysis*

For all tests, the Type I error rate was .05. Because the dependent variable, predictive bias, was dichotomous (0,1), logistic regression was used (Hosmer & Lemeshow, 2000). This procedure was chosen for two reasons: (a) the relation between the dependent variable and the independent variable is non-linear, due to the bounded nature of the dependent variable (0,1); this violates the linearity assumption of ANOVA and linear regression and (b) the error of

residuals is non-normal for binary variables, which also violates the assumption of homoscedasticity in linear models. Logistic regression accommodates the bounded nature of the dependent variable and does not depend on the homoscedasticity assumption (Hosmer & Lemeshow, 2000). The following logistic model was employed to test Hypotheses 1-4:

$$P(y_i = 1 | \mathbf{v}) = \{1 + \text{Exp}[-1(b_0 + b_1 v_1 + b_2 v_2 + b_3 v_3 + b_4 v_4 + b_5 v_5 + b_6 v_6)]\}^{-1}, \quad (16)$$

where  $P(y_i = 1 | \mathbf{v})$  is the probability of detecting predictive bias given  $\mathbf{v}$ ,  $\mathbf{v} = [v_1, v_2, v_3, v_4, v_5, v_6]$ ,  $v_1$  = predictor difference,  $v_2$  = criterion difference,  $v_3$  = validity coefficient,  $v_4$  = sample size,  $v_5$  = percentage of DIF, and  $v_6$  = effect size of DIF.

This main effects model was used to determine the extent to which the hypothesized variables influenced the occurrence of predictive bias against the reference group and the focal group. Hypothesis 1 was evaluated by the presence of a significant logit coefficient for the percentage of DIF variable ( $v_5$ ). Hypothesis 2 was examined by the existence of a significant coefficient for the effect size of DIF variable ( $v_6$ ). Hypothesis 3 was tested by the presence of a significant logit coefficient for the sample size variable ( $v_4$ ). Hypotheses 4A and 4B were examined by the presence of a statistically significant logit coefficient for the predictor ( $v_1$ ) and criterion difference variables ( $v_2$ ), respectively. In an exploratory analysis, interactions among the independent variables were also modeled in the logistic regression analysis.

## RESULTS

### *Summary Tables*

Results of the simulations are summarized in Tables 4-10. In the tables, each value was calculated as the proportion of times that predictive bias was detected out of 500 replications. Predictive bias was indicated in Equation 7 by a significant difference in subgroup intercepts or subgroup slopes (i.e., the rejection of the null hypothesis that either  $\beta_2 = 0$  or  $\beta_3 = 0$ ) in each replication of a condition. Within each table, the rate of detecting predictive bias against the reference group is summarized in Part A, and the rate of detecting predictive bias against the focal group is summarized in Part B.

*Predictive bias against the reference group.* Although there was no DIF in any condition in Table 4A, predictive bias occurred against the reference group at a rate greater than chance when there was a difference on the criterion. For example, when there was a mean subgroup difference of .35 on the criterion and no mean subgroup difference on the predictor, predictive bias against the reference group was detected from approximately 18% to 76% of the time when the validity coefficient was .3. Predictive bias also occurred against the reference group when there were approximately equal mean subgroup differences on both the predictor and the criterion. When there were mean subgroup differences on the predictor and criterion of .33 and .35, respectively, the rates of detecting predictive bias against the reference group were from approximately 14% to 47% when the validity coefficient was .3.

*Predictive bias against the focal group.* In Table 4B, when there was no DIF in any condition, predictive bias against the focal group occurred at a rate greater than chance when

there was a difference in subgroup means on the predictor. For example, when there was a mean subgroup difference of one *SD* on the predictor and no mean subgroup difference on the criterion, predictive bias against the focal group was detected from approximately 18% to 86% of the time, varying as a function of the validity coefficient (.3 and .5) and subgroup sample size (35, 70, and 105). However, when there were mean subgroup differences on both the predictor and criterion, predictive bias against the focal group occurred less than one would expect. For instance, when there were approximately equal mean subgroup differences on both the predictor and criterion of .33 and .35, respectively, predictive bias against the focal group was detected from approximately 1% to 3% of the time under varying conditions of validity coefficients and sample sizes. In summary, predictive bias was detected against both subgroups when DIF (or measurement bias) was not present.

*Test of Hypotheses 1-4.* The logistic regression model in Equation 16 was used to evaluate Hypotheses 1-4. Six independent variables were manipulated: (a) four levels of predictor difference, (b) two levels of criterion difference, (c) two levels of validity coefficient, (d) three levels of sample size, (e) three levels of percentage of DIF, and (f) three levels of DIF effect size. When these variables were completely crossed, there were 432 conditions. Each condition was replicated 500 times for a total of 216,000 observations. Two criteria were used to select an adequate sample of data to analyze: sample size and number of events (in this case, the number of times predictive bias was detected) per parameter in the logistic regression model. Although sample size considerations were important in the analysis of the simulated data, past research showed that the number of events per parameter in the logistic model has serious implications for stable parameter estimates and hypothesis testing (Hosmer & Lemeshow, 2000).

Table 4.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and No Differential Item Functioning (% of DIF = 0, Effect Size of DIF = 0)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.04	.04	.18	.38
	70	.05	.06	.55	.63
	105	.07	.05	.76	.77
.33	35	.02	.02	.14	.15
	70	.02	.01	.31	.25
	105	.05	.03	.47	.35
.66	35	.04	.02	.06	.07
	70	.05	.04	.14	.10
	105	.05	.05	.22	.10
1.0	35	.02	.06	.05	.05
	70	.05	.07	.08	.08
	105	.05	.08	.10	.11

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.03	.05	.01	.03
	70	.04	.06	.03	.03
	105	.07	.06	.02	.02
.33	35	.06	.12	.01	.02
	70	.08	.21	.03	.01
	105	.13	.23	.02	.03
.66	35	.10	.27	.00	.03
	70	.19	.49	.01	.02
	105	.27	.63	.02	.02
1.0	35	.18	.48	.02	.07
	70	.32	.69	.02	.11
	105	.45	.86	.02	.14

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Table 5.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor and Criterion,  
 Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential  
 Item Functioning (% of DIF = 15, Effect Size of DIF = .3)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor difference	Sample size	Validity coefficient		Validity coefficient	
		.3	.5	.3	.5
.00	35	.03	.05	.21	.32
	70	.04	.04	.49	.58
	105	.04	.04	.71	.81
.33	35	.03	.03	.10	.16
	70	.04	.03	.31	.23
	105	.04	.03	.44	.30
.66	35	.03	.05	.07	.07
	70	.03	.05	.14	.07
	105	.03	.04	.21	.08
1.0	35	.03	.07	.06	.08
	70	.04	.07	.08	.07
	105	.04	.11	.10	.09

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.05	.05	.01	.02
	70	.05	.06	.02	.03
	105	.05	.07	.03	.02
.33	35	.06	.15	.01	.02
	70	.11	.22	.01	.03
	105	.14	.27	.02	.02
.66	35	.12	.28	.02	.03
	70	.21	.51	.02	.02
	105	.28	.68	.02	.05
1.0	35	.15	.46	.02	.07
	70	.30	.76	.02	.13
	105	.46	.84	.02	.14

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Table 6.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning  
(% of DIF = 15, Effect Size of DIF = .6)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.03	.05	.21	.33
	70	.04	.06	.49	.55
	105	.05	.03	.67	.71
.33	35	.03	.04	.09	.12
	70	.04	.04	.28	.20
	105	.04	.04	.43	.28
.66	35	.03	.05	.07	.06
	70	.04	.05	.13	.07
	105	.05	.06	.19	.09
1.0	35	.04	.07	.03	.07
	70	.05	.10	.09	.09
	105	.06	.10	.13	.09

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.04	.05	.02	.03
	70	.04	.06	.03	.02
	105	.07	.05	.01	.01
.33	35	.05	.18	.02	.01
	70	.10	.24	.02	.02
	105	.15	.37	.02	.02
.66	35	.13	.28	.02	.04
	70	.21	.57	.01	.04
	105	.30	.71	.01	.04
1.0	35	.17	.46	.01	.08
	70	.35	.75	.02	.13
	105	.45	.85	.01	.17

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Table 7.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 15, Effect Size of DIF = .9)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor difference	Sample size	Validity coefficient		Validity coefficient	
		.3	.5	.3	.5
.00	35	.04	.06	.17	.27
	70	.04	.04	.43	.52
	105	.04	.03	.66	.67
.33	35	.03	.04	.11	.13
	70	.04	.05	.27	.16
	105	.04	.04	.37	.27
.66	35	.03	.04	.07	.06
	70	.04	.06	.15	.08
	105	.04	.09	.18	.09
1.0	35	.03	.06	.05	.06
	70	.04	.11	.08	.08
	105	.06	.11	.10	.12

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.04	.05	.01	.02
	70	.07	.07	.03	.02
	105	.07	.06	.01	.02
.33	35	.06	.17	.03	.02
	70	.17	.32	.02	.01
	105	.16	.36	.02	.01
.66	35	.14	.30	.01	.03
	70	.21	.60	.02	.04
	105	.33	.69	.01	.04
1.0	35	.18	.51	.03	.11
	70	.32	.74	.02	.12
	105	.47	.85	.01	.18

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Table 8.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 30, Effect Size of DIF = .3)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor difference	Sample size	Validity coefficient		Validity coefficient	
		.3	.5	.3	.5
.00	35	.03	.04	.18	.27
	70	.03	.05	.46	.53
	105	.06	.02	.67	.69
.33	35	.04	.03	.10	.14
	70	.04	.03	.27	.18
	105	.03	.04	.41	.26
.66	35	.03	.05	.04	.06
	70	.04	.04	.12	.07
	105	.03	.06	.17	.08
1.0	35	.02	.05	.03	.05
	70	.05	.08	.07	.11
	105	.05	.12	.09	.12

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.05	.07	.02	.01
	70	.06	.07	.02	.02
	105	.07	.06	.01	.01
.33	35	.06	.15	.01	.02
	70	.13	.25	.03	.01
	105	.15	.37	.02	.02
.66	35	.10	.33	.02	.04
	70	.22	.54	.02	.02
	105	.32	.70	.01	.03
1.0	35	.14	.54	.02	.10
	70	.30	.79	.03	.14
	105	.53	.84	.02	.18

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Table 9.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning  
(% of DIF = 30, Effect Size of DIF = .6)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor difference	Sample size	Validity coefficient		Validity coefficient	
		.3	.5	.3	.5
.00	35	.03	.04	.16	.24
	70	.04	.04	.39	.42
	105	.04	.02	.61	.55
.33	35	.02	.04	.08	.10
	70	.03	.03	.21	.15
	105	.03	.03	.33	.16
.66	35	.03	.04	.04	.07
	70	.04	.06	.12	.08
	105	.03	.07	.13	.09
1.0	35	.03	.08	.04	.06
	70	.05	.07	.06	.10
	105	.05	.13	.09	.13

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.05	.08	.01	.03
	70	.06	.11	.03	.02
	105	.07	.11	.02	.03
.33	35	.06	.19	.01	.03
	70	.14	.32	.02	.02
	105	.16	.48	.01	.01
.66	35	.12	.35	.02	.04
	70	.26	.62	.02	.05
	105	.35	.79	.02	.08
1.0	35	.16	.53	.01	.10
	70	.38	.83	.04	.18
	105	.54	.85	.03	.26

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Table 10.

Predictive Bias as a Function of a Mean Subgroup Difference on the Predictor, Mean Subgroup Difference on the Criterion, Validity Coefficient, Sample Size, and Differential Item Functioning (% of DIF = 30, Effect Size of DIF = .9)

(A) Reference group		Criterion difference			
		.00		.35	
Predictor difference	Sample size	Validity coefficient		Validity coefficient	
		.3	.5	.3	.5
.00	35	.03	.04	.13	.22
	70	.03	.04	.36	.32
	105	.03	.04	.54	.46
.33	35	.01	.04	.10	.09
	70	.03	.07	.16	.10
	105	.05	.07	.31	.14
.66	35	.03	.05	.05	.05
	70	.04	.08	.11	.06
	105	.06	.10	.15	.07
1.0	35	.04	.10	.05	.07
	70	.08	.15	.08	.13
	105	.08	.16	.10	.17

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

(B) Focal group		Criterion difference			
		.00		.35	
Predictor	Sample	Validity coefficient		Validity coefficient	
difference	size	.3	.5	.3	.5
.00	35	.06	.10	.02	.01
	70	.08	.12	.02	.02
	105	.11	.19	.02	.02
.33	35	.09	.26	.01	.02
	70	.16	.41	.01	.02
	105	.27	.56	.01	.01
.66	35	.13	.37	.02	.06
	70	.27	.69	.02	.06
	105	.40	.80	.00	.09
1.0	35	.16	.59	.03	.14
	70	.37	.77	.03	.25
	105	.52	.83	.03	.26

*Note.* Each cell represents the proportion of times that predictive bias was detected out of 500 replications.

Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) showed that the occurrence of the least frequent outcome on the dependent variable adversely influences variance estimates of the logit coefficients, which subsequently influences the Wald test of significance. Thus, Hosmer and Lemeshow (2000) and Peduzzi et al. (1996) recommended that samples have approximately 10 or more events per parameter in the logistic regression model. Because there were seven parameters in the model (i.e., one constant and one for each of the six independent variables in Equation 16), data were randomly sampled so that the dependent variable (predictive bias against the reference or focal group) had at least 70 or more events of predictive bias. With all conditions completely crossed and meeting the events per parameter requirement, 864 observations from the 216,000 observations were sampled with 110 events of predictive bias against the reference group and 139 events of predictive bias against the focal group. The summary of the logistic regression analysis of predictive bias against the reference group is presented first, which is followed by a summary of the results of the logistic regression analysis of predictive bias against the focal group.

#### *Reference Group Analysis*

Results of the logistic regression model (Equation 16) with predictive bias against the reference group as the dependent variable are presented in Table 11. For the overall model,  $\chi^2(6) = 127.36, p < .01$ , accounting for approximately 14% of the variance in detecting predictive bias against the reference group.

*Percentage of DIF.* Hypothesis 1 stated that as the percentage of DIF increases, the rate of detecting predictive bias increases. Support for this hypothesis was not found. The logit

coefficient in Equation 16 for the percentage of DIF variable was zero,  $\chi^2(1) = 0, p > .05$ . This is also consistent with the results that are summarized in Part A of Tables 4-10 for the reference group. As DIF increased from zero in Table 4A to 30% in Table 10A, there was no substantial increase in the proportion of times that predictive bias was detected against the reference group. For example, when there were no mean differences on the predictor and criterion, detection of predictive bias against the reference group was from approximate 2% to 7% across conditions.

Table 11.

Summary of the Logistic Regression Analysis for Variables Predicting  
 Predictive Bias Against the Reference Group ( $N = 864$ )

Variable	$b$	$SE(b)$	$\chi^2$
Constant	-2.56	.16	241.54**
Predictor difference	-.62	.12	26.65**
Criterion difference	1.13	.15	56.09**
Validity coefficient	-.03	.11	0.05
Sample size	.52	.12	19.77**
Percentage of DIF	.00	.11	0.00
Effect size of DIF	.08	.11	0.46

Note. \*  $p \leq .05$ , \*\*  $p \leq .01$ ; Cox and Snell  $R^2 = .14, \chi^2(6) = 127.36, p < .01$

*Effect size of DIF.* Hypothesis 2 stated that as the effect size of DIF increases, the rate of detecting predictive bias increases. There was no evidence in support of this hypothesis. The logit coefficient in Equation 16 for the effect size of DIF variable was  $b_6 = .08$ ,  $\chi^2(1) = 0.464$ ,  $p > .05$ . The null results for Hypotheses 1 and 2 were perhaps due to the fact that DIF was created to be against the focal group and not the reference group, thus it had no substantial influence on detecting predictive bias against the reference group.

*Sample size.* Hypothesis 3 stated that as the sample size increases, the rate of detecting predictive bias increases. Evidence was found to support this notion. The logit coefficient in Equation 16 for sample size was  $b_4 = .52$ ,  $\chi^2(1) = 19.77$ ,  $p < .01$ . The results are illustrated in Figure 12. The figure shows that as subgroup sample size increased from 35 per group to 105 per group, the rate of detecting predictive bias against the reference group increased from approximately 7% to 19%.

*Predictor difference.* Hypothesis 4A stated that as the mean subgroup difference on the predictor increases, predictive bias increases. There was no evidence in support of this hypothesis. The mean subgroup difference on the predictor had an influence in the detection of predictive bias against the reference group,  $b_1 = -.62$ ,  $\chi^2(1) = 26.65$ ,  $p < .01$ , but the results were not in the hypothesized direction. This suggests that as the mean subgroup difference on the predictor increases, the rate of detecting predictive bias against the reference group decreases.

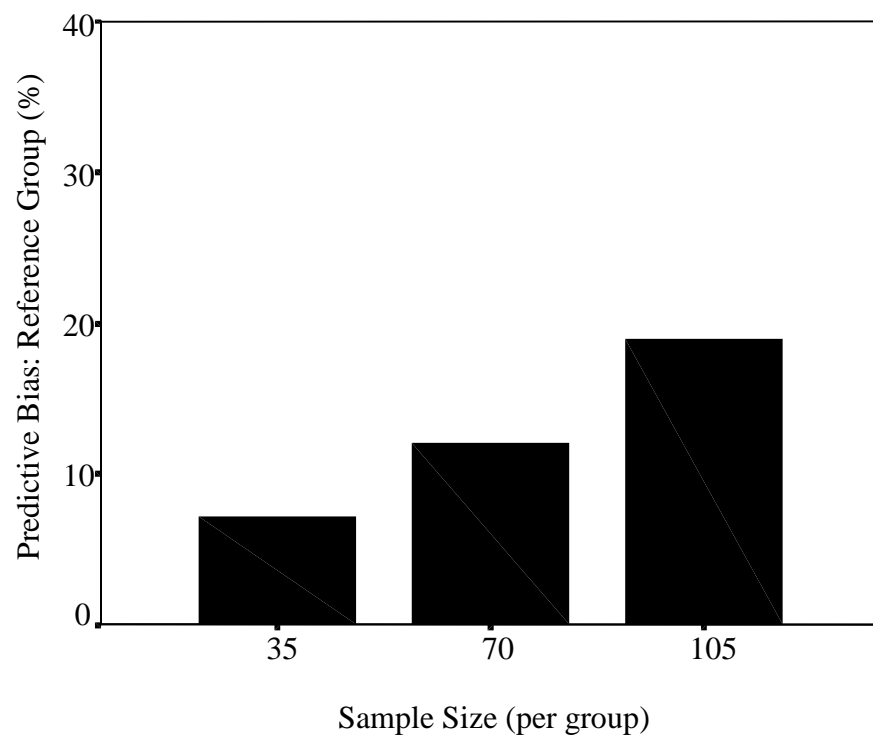


Figure 12. Predictive bias as a function of sample size: Reference group.

*Criterion difference.* Hypothesis 4B stated that as the mean subgroup difference on the criterion increases, predictive bias increases. Some evidence was found in support of this hypothesis. In Equation 16, the coefficient for the criterion difference variable was  $b_2 = 1.134$ ,  $\chi^2(1) = 56.09, p < .01$ . As shown in Figure 13, when the difference on the criterion increased from zero to  $.35 SD$  in favor of the reference group, the detection of predictive bias against the reference group increased from approximately 3% to 22%.

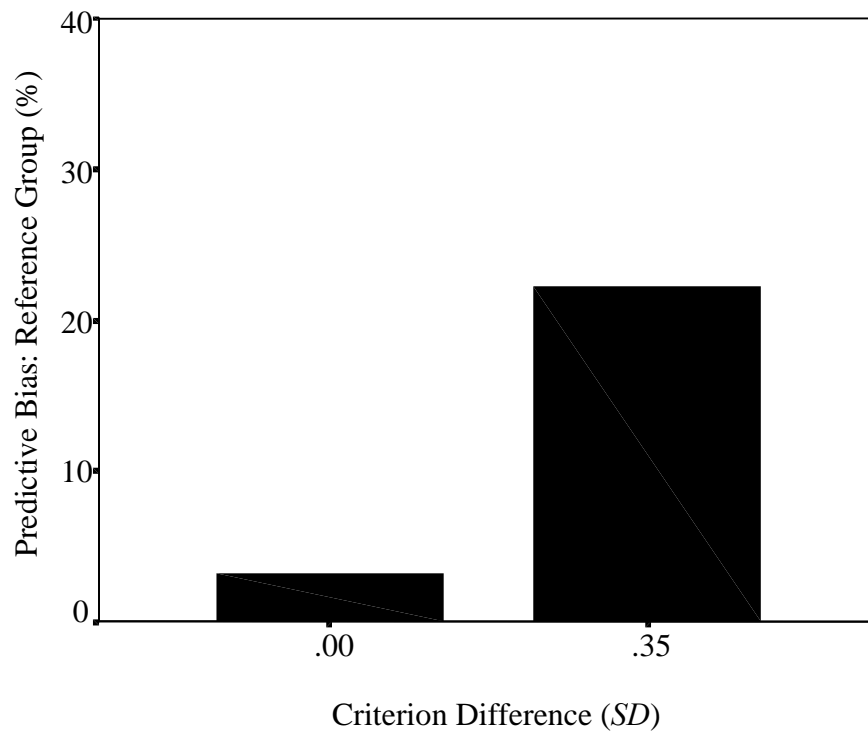


Figure 13. Predictive bias as a function of criterion difference: Reference group.

### *Focal Group Analysis*

Results of the logistic regression analysis (Equation 16) for predictive bias against the focal group are presented in Table 12. For the overall model,  $\chi^2(6) = 248.44, p < .01$ , accounting for approximately 25% of the variance in the incidence of predictive bias against the focal group.

*Percentage of DIF.* Hypothesis 1 stated that as the percentage of DIF increases, the rate of detecting predictive bias increases. Evidence in support of this hypothesis was not found. The coefficient for the percentage of DIF variable was  $b_5 = .14, \chi^2(1) = 1.52, p > .05$ .

Table 12.

Summary of the Logistic Regression Analysis for Variables Predicting  
Predictive Bias Against the Focal Group ( $N = 864$ )

Variable	$b$	$SE(b)$	$\chi^2$
Constant	-2.56	.18	220.53**
Predictor difference	.99	.13	59.64**
Criterion difference	-1.38	.15	83.10**
Validity coefficient	.79	.12	41.52**
Sample size	.66	.12	30.60**
Percentage of DIF	.14	.11	1.52
Effect size of DIF	.11	.11	0.92

Note. \*  $p \leq .05$ , \*\*  $p \leq .01$ ; Cox and Snell  $R^2 = .25, \chi^2(6) = 248.44, p < .01$

*Effect size of DIF.* Hypothesis 2 stated that as the effect size of DIF increases, the rate of predictive bias increases. There was no support for this hypothesis. The logit coefficient in Equation 16 for the effect size of DIF variable was  $b_6 = .11$ ,  $\chi^2(1) = 0.92$ ,  $p > .05$ . Thus, there was no evidence that DIF is related to predictive bias.

*Sample size.* Hypothesis 3 stated that as the sample size increases, the detection of predictive bias increases. In contrast to Hypotheses 1 and 2, evidence was found to support this view. In Equation 16, the logit coefficient for sample size was  $b_4 = .66$ ,  $\chi^2(1) = 30.60$ ,  $p < .01$ . As shown in Figure 14, when subgroup sample size was low (35), the detection of predictive bias against the focal group was approximately 9%, but when the subgroup sample size was high (105), predictive bias against the focal group increased to approximately 24%.

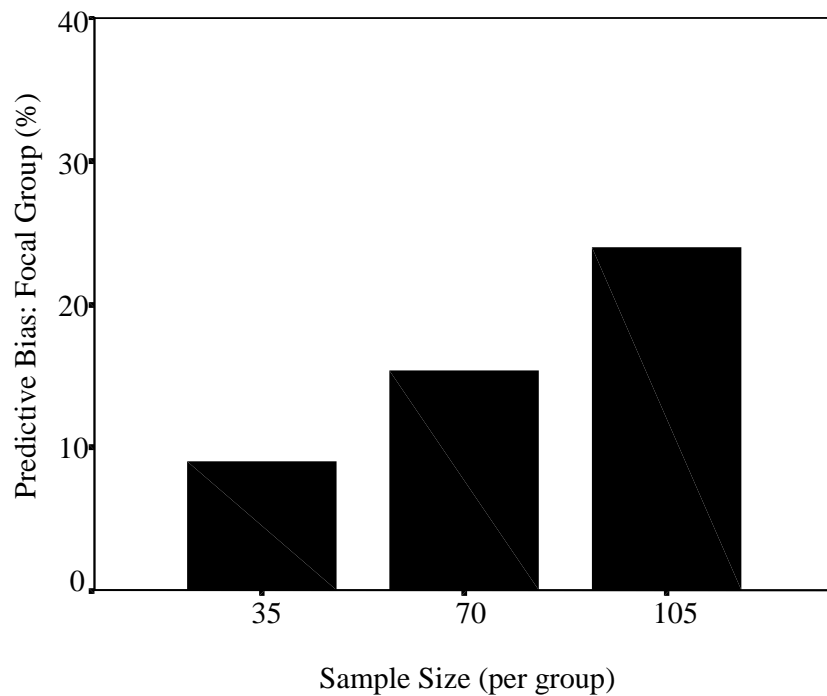


Figure 14. Predictive bias as a function of sample size: Focal group.

*Predictor difference.* Hypothesis 4A stated that as the mean difference on the predictor increases, the detection of predictive bias increases. There was support for this hypothesis. The coefficient for the predictor difference variable ( $b_1$ ) was .99,  $\chi^2(1) = 59.64, p < .01$ . As shown in Figure 15, when the mean subgroup difference on the predictor increased from zero to one *SD* difference, the rate of detecting predictive bias against the focal group increased from approximately 5% to 30%.

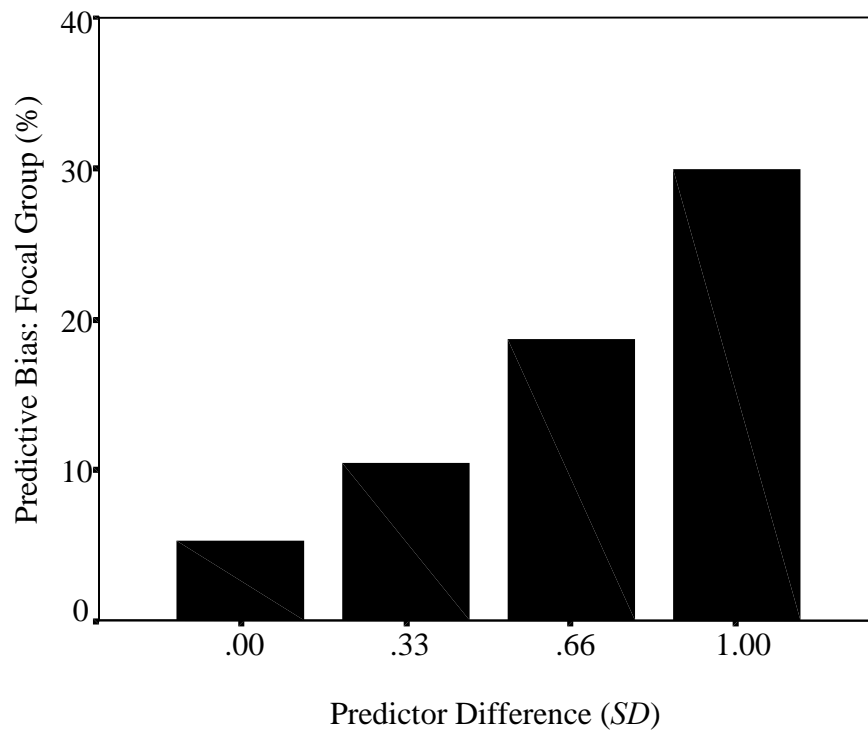


Figure 15. Predictive bias as a function of predictor difference: Focal group.

*Criterion difference.* Hypothesis 4B stated that as the mean difference on the criterion increases, the detection of predictive bias increases. There was no support for this hypothesis, for the results were not in the hypothesized direction. The mean subgroup difference on the criterion had an influence on the detection of predictive bias against the focal group,  $b_2 = -1.38$ ,  $\chi^2(1) = 83.10$ ,  $p < .01$ . This indicates that as the mean subgroup difference on the criterion increased, the rate of detecting predictive bias against the focal group decreased.

Across both analyses of predictive bias against the reference group and the focal group, DIF (i.e., percentage of DIF and effect size of DIF) had no statistically significant influence on predictive bias, which was contrary to Hypotheses 1 and 2. With over 800 observations and over 10 events per parameter in each analysis, there was sufficient power to detect the effect of DIF on predictive bias. Sample size and ability differences (either on the predictor or criterion) had a significant influence on the rate of detecting predictive bias for both reference and focal groups. This provided some support for Hypotheses 3 and 4. However, further analyses were warranted for the reasons described below.

### *Exploratory Analyses*

Exploratory analyses were also conducted. First, because no moderating hypotheses were given, main effects were modeled in the previous analyses without considering interactions. Thus, it was plausible that the main effects model (Equation 16) in this study was incorrectly specified if higher order interactions were present in the data but were undetected. This warranted a logistic regression model with moderating effects included. Second, a replication of the results using another sample from the simulated data set, a fully specified logistic regression

model (i.e., all interactions included), and a larger number of observations would demonstrate the robustness of the findings from the previous analyses.

Because of the number of the independent variables (six) in this study, 63 parameters were needed to create a fully saturated model with all 2-, 3-, 4-, 5-, and 6-way interactions. This model is an extension of Equation 16 with all higher-order interactions. Thus, a larger sample and a larger number of events of predictive bias were required for both the reference and focal group analyses. From the 215,136 observations remaining after the first set of logistic regression analyses were completed, the exploratory analyses were done with a sample of 4,752 observations with 627 occurrences of predictive bias against the reference group and 781 occurrences of predictive bias against the focal group. Because the conditions were totally crossed, the reciprocal of the correlation matrix of predictors was an identity matrix, indicating that multicollinearity of the predictors was not an issue.

*Exploratory analysis: Reference group.* Results of the logistic regression analysis for predictive bias against the reference group using the full model were similar to results of the main effects model with some exceptions. For the overall model,  $\chi^2(62) = 968.93, p < .01$ , accounting for 18% of the variance in detecting predictive bias against the reference group. Similar to the main effects model above, percentage of DIF had no influence on the rate of detecting predictive bias against the reference group,  $b_5 = 0, \chi^2(1) = 0, p > .05$ . There was also no evidence in support of Hypothesis 2: The effect size of DIF had no influence on the probability of detecting predictive bias against the reference group,  $b_6 = -.01, \chi^2(1) = .01, p > .05$ .

Consistent with the main effects model presented earlier, the logit coefficients representing the sample size variable,  $b_4 = .29$ ,  $\chi^2(1) = 15.33$ ,  $p < .01$ , and the criterion difference variable,  $b_2 = .93$ ,  $\chi^2(1) = 154.97$ ,  $p < .01$ , provided support for Hypotheses 3 and 4B, respectively, replicating the results for the full model. The results for the predictor difference variable were also similar to the first analysis,  $b_1 = -.27$ ,  $\chi^2(1) = 14.07$ ,  $p < .01$ . The interpretations for all effects were similar to the results of the main effects model and were not presented. However, the effects above pertaining to the predictor and criterion difference variables were considered to be conditional due to the presence of a statistically significant interaction described below.

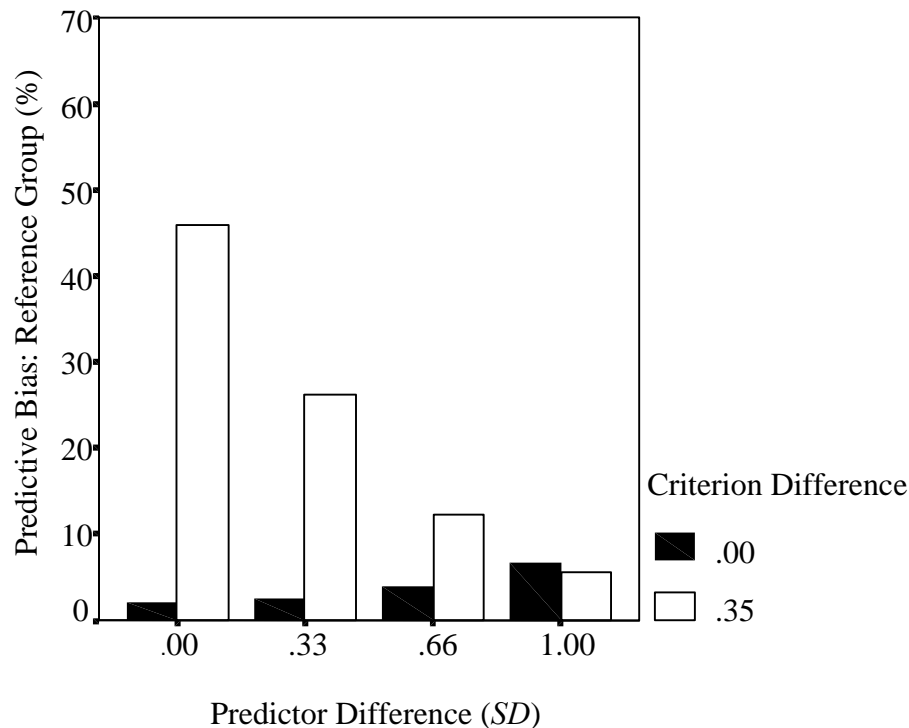


Figure 16. Predictive bias as a function of predictor and criterion difference: Reference group.

There was a two-way interaction between the predictor difference and criterion difference variables in detecting predictive bias against the reference group,  $b_7 = -.78$ ,  $\chi^2(1) = 117.87$ ,  $p < .01$ . The predicted probabilities are illustrated in Figure 16. As shown in the figure, when there was no difference on the criterion, the probability of detecting predictive bias against the reference group slightly increased from approximately 2% to 7% as the mean subgroup difference on the predictor increased from no difference to one *SD* difference, respectively. However, when there was a mean subgroup difference on the criterion of .35, the probability of detecting predictive bias against the reference group decreased from approximately 47% to 6% as the mean subgroup difference on the predictor increased from no difference to one *SD*, respectively. In other words, the mean subgroup difference on the criterion moderated the relation between the mean subgroup difference on the predictor and the probability of detecting predictive bias. The interpretation is consistent with the results in Table 4A.

*Exploratory analysis: Focal group.* Results of the logistic regression analysis for detecting predictive bias against the focal group also replicated the results of the main effects model. For the overall model,  $\chi^2(62) = 1347.72$ ,  $p < .01$ , accounting for 25% of the variance in detecting predictive bias against the focal group. Similar to the previous analysis, the percentage of DIF in the full model had no influence on the probability of detecting predictive bias against the focal group,  $b_5 = .03$ ,  $\chi^2(1) = .19$ ,  $p > .05$ . There was also no evidence in support of Hypothesis 2; the effect size of DIF had no influence on the rate of detecting predictive bias,  $b_6 = .07$ ,  $\chi^2(1) = .76$ ,  $p > .05$ . Also consistent with the main effects model, the coefficient for the sample size variable in the full model was  $b_4 = .15$ ,  $\chi^2(1) = 3.84$ ,  $p < .05$ , providing support for

Hypotheses 3. The predictor difference variable in the full model was  $b_1 = .68$ ,  $\chi^2(1) = 88.80$ ,  $p < .01$ , which is also consistent with the *a priori* hypothesis (Hypothesis 4A) and the results for the main effects model presented earlier. The results for the criterion difference variable were also similar to the first analysis,  $b_2 = -1.19$ ,  $\chi^2(1) = 233.82$ ,  $p < .01$ , but both predictor difference and criterion difference variables also interacted, thus the effects were considered as conditional.

There was a significant two-way interaction of the predictor difference and criterion difference variables in detecting predictive bias against the focal group,  $b_7 = -.41$ ,  $\chi^2(1) = 32.11$ ,  $p < .01$ . The predicted probabilities are illustrated in Figure 17. As shown in the figure, when there was no difference on the criterion, the probability of detecting predictive bias against the focal group increased from approximately 7% to 55% as the mean subgroup difference on the predictor increased from no difference to one *SD* difference. In contrast, when there was a mean subgroup difference on the criterion of .35, the probability of detecting predictive bias against the focal group increased slightly from approximately 3% to 7% as the mean subgroup difference on the predictor increased from no difference to a one *SD* difference. Thus, the mean subgroup difference on the criterion moderated the relation between the mean subgroup difference on the predictor and the probability of detecting predictive bias against the focal group. The interpretation of the interaction is consistent with the results in Table 4B.

There was also a three-way interaction among the predictor difference, criterion difference, and percentage of DIF variables in the probability of detecting predictive bias against the focal group,  $b_{24} = .20$ ,  $\chi^2(1) = 7.71$ ,  $p < .01$ . The three-way interaction is illustrated in Figures 18-20.

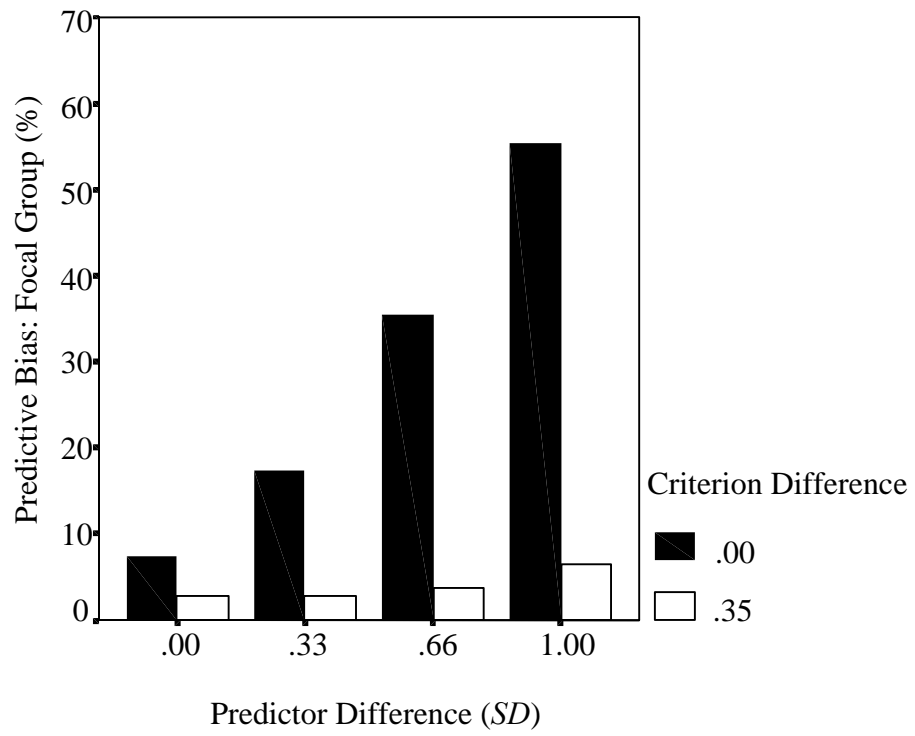


Figure 17. Predictive bias as a function of predictor and criterion difference: Focal group.

In Figure 18, when DIF was not present and there was no mean subgroup difference on the criterion, the rate of detecting predictive bias against the focal group increased as the mean subgroup difference on the predictor increased. This is consistent with the summary of results in Table 4B, which suggests that predictive bias can exist without DIF being present. However, when there was a mean subgroup difference on the criterion of  $.35 SD$ , the rate of detecting predictive against the focal group remained at approximately 3% when the mean subgroup difference on the predictor increased from zero to one  $SD$ .

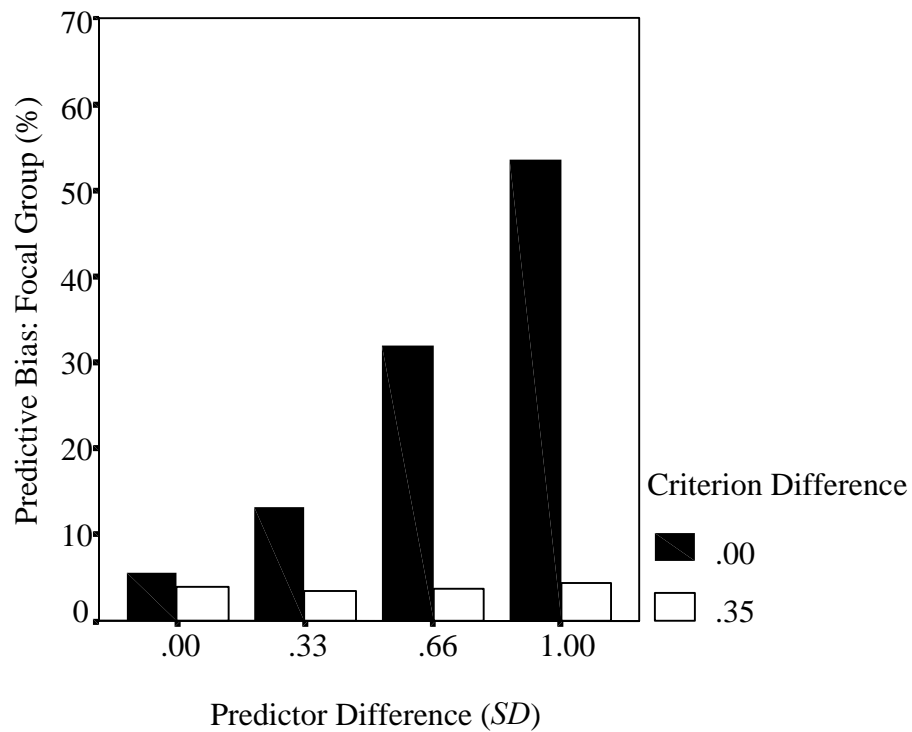


Figure 18. Predictive bias as a function of predictor and criterion difference when percentage of DIF = 0: Focal group.

In Figure 19, when DIF occurred at a rate of 15% against the focal group and there was no mean subgroup difference on the criterion, the rate of detecting predictive bias against the focal group increased as the mean subgroup difference on the predictor increased. In contrast, when DIF occurred at a rate of 15% against the focal group and there was a mean subgroup difference on the criterion (.35), there was a slight increase in the rate of detecting predictive bias against the focal group as the mean subgroup difference on the predictor increased. However, the rate of detecting predictive bias against the focal group did not exceed what would be expected by chance. This suggests that when there is a difference on the criterion in favor of the reference group and DIF against the focal group is present, the detection of predictive bias against the focal group is not likely to occur.

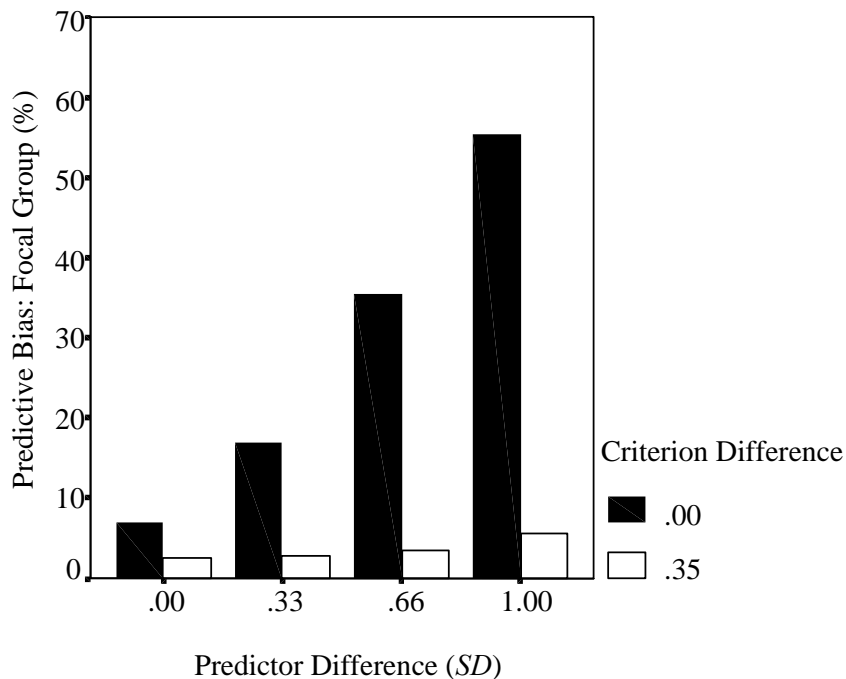


Figure 19. Predictive bias as a function of predictor and criterion difference when percentage of DIF = 15: Focal group.

When there was a mean subgroup difference on the criterion (.35) and DIF occurred at a rate of 30%, the probability of detecting predictive bias against the focal group increased slightly as the mean subgroup difference on the predictor increased. As shown in Figure 20, when there was a mean criterion difference of .35, the rate of detecting predictive bias against the focal group slightly increased as the mean subgroup difference on the predictor increased. For instance, when there was no difference on the predictor, the rate of detecting predictive bias was approximately 2%, but when there was a mean subgroup difference on the predictor of one *SD*, the rate of detecting predictive bias was approximately 10%. This provides evidence in support of the notion that DIF, if present, has a moderating effect on predictive bias against the focal group under conditions that have been reported in the literature.

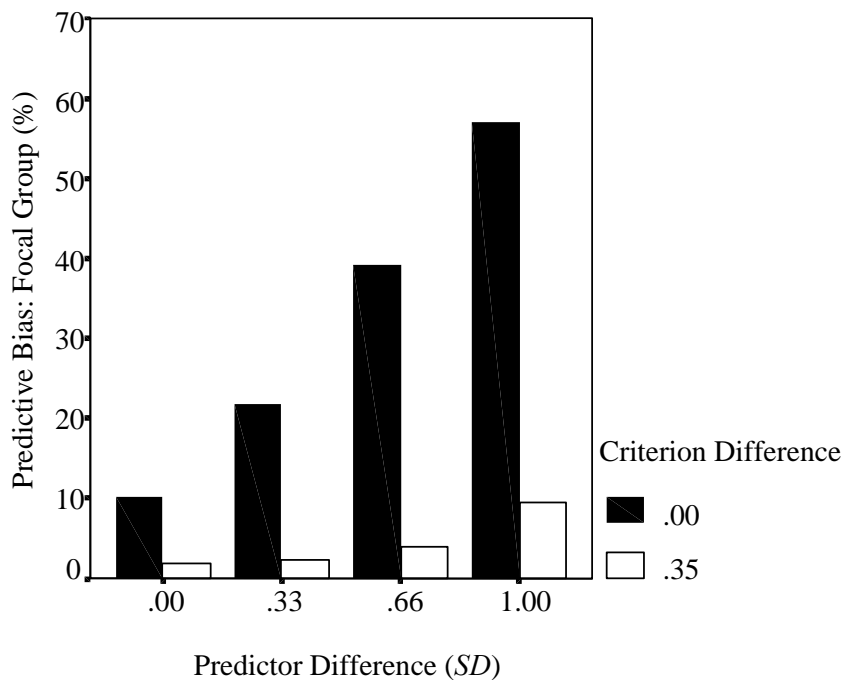


Figure 20. Predictive bias as a function of predictor and criterion difference when percentage of DIF = 30: Focal group.

## DISCUSSION

### *Summary of Findings*

The purpose of this research was to investigate the relation between measurement bias at the item level (DIF) and predictive bias at the test score level. Specifically, does DIF have an effect on predictive bias at the test score level? Because some researchers have argued that measurement bias and predictive bias are mutually supportive, i.e., the absence (or presence) of one type of bias is evidence of the absence (or presence) of the other type of bias, this study sought to answer two ancillary, but very important, questions about the relation between predictive bias and DIF. First, can predictive bias against a subgroup exist when DIF is not present? Second, can DIF against a subgroup exist when predictive bias is not present?

The summary of the simulations in Tables 4-10 provides some evidence to support the notion that predictive bias against a subgroup can exist when DIF is not present. Probabilities in Table 4 are indicative of the existence of predictive bias against both reference and focal groups when DIF does not exist. For example, Table 4A shows that the probability of detecting predictive bias against the reference group varied from approximately 2% to 77% for different sample sizes, mean subgroup differences on the predictor/criterion, and validity coefficients. Table 4B also shows that the chances of detecting predictive bias against the focal group varied from approximately 0% to 86% under the conditions examined in this study. This suggests that predictive bias against both reference and focal groups may occur independently of measurement bias at the item level. Formal hypothesis testing was also used to investigate the relation between DIF and predictive bias.

*Percentage of DIF and effect size of DIF.* The results of the logistic regression analyses are consistent with the notion that there is no direct relation between DIF and predictive bias. In the main effects model that examined predictive bias against both reference and focal groups, the percentage of DIF and the effect size of DIF were not significant predictors of the probability of detecting predictive bias. Because the percentage of DIF and the effect size of DIF manipulations made items more difficult for the focal group holding ability constant, it is reasonable to believe that predictive bias would occur against the focal group but probably would not occur against the reference group. However, the results showed that DIF had no influence on the rate of detecting predictive bias against either subgroup. In the exploratory analyses, when there was more than enough power in terms of the number of observations sampled (over 4,500) to detect an effect, DIF was not related to predictive bias. As an explanation for these findings, it is plausible that the null result is due to the fact that there was no measurement bias manipulation in the same direction on the criterion. Future research could investigate the degree to which measurement bias on both the predictor and criterion influences the rate of detecting predictive bias under similar conditions presented in this investigation. At this point, the evidence in this study is contrary to what some researchers posit about the mutually supportive relation between predictive bias and DIF (Hunter & Schmidt, 2000; Hunter et al., 1984).

*Sample size effects.* Results also showed that the detection of predictive bias against both the focal group and reference group was influenced by sample size, thus providing support for Hypothesis 3 across both subgroup analyses. As described in Scenario 2, sample size has an influence on the standard error term; specifically, the standard error of the regression coefficient decreases as the sample size increases, as implied by Equation 14. So, this finding is not surprising given what is known about the relation between the standard error and sample size.

However, because it is not central to the questions asked in this study, this finding is no longer discussed.

*Predictor and criterion effects.* Evidence across both reference and focal group analyses provides mixed support for the position that when differences increase on the predictor or criterion, the rate of detecting predictive bias increases (Hypothesis 4A and 4B). The results of the main effects model for the reference group showed that when there was a mean subgroup difference on the criterion, predictive bias against the reference group increased. On average, when there was a mean subgroup difference on the criterion of .35 in favor of the reference group, predictive bias against the reference group occurred approximately 22% of the time. Note the striking similarity to the 24% estimate of predictive bias found in over 1,100 studies reported by Bartlett et al. (1978); these studies also revealed that the bias was against non-minorities. Thus, the selection of an unbiased criterion has implications for predictive bias. The results also showed, however, that as the mean subgroup difference on the predictor increased, predictive bias against the reference group decreased.

In the main effects model for the focal group, results showed that as the mean subgroup difference on the predictor increased, predictive bias against the focal group increased: Detection rates were from approximately 5% (no difference) to approximately 30% (one *SD* difference). However, the results of the main effects analyses should be viewed as conditional effects in the presence of interactions in the fully specified model. The results of the exploratory analyses showed that the influence of the mean predictor difference on the probability of detecting predictive bias was moderated by the mean subgroup difference on the criterion.

*Interaction effects.* The results of the exploratory analyses suggest that the effects of ability differences on predictive bias were not direct as hypothesized. These analyses across both reference and focal groups showed that there was a significant two-way interaction between mean subgroup predictor differences and mean subgroup criterion differences in detecting predictive bias. However, the effects of the two-way interaction of the predictor and criterion differences on predictive bias were not the same for the two subgroups.

The results of the reference group analysis showed that when there was a mean subgroup difference on the criterion and no difference on the predictor, the rate of detecting predictive bias against the reference group was moderately high (approximately 46%), but decreased dramatically to the chance level (5%) as the mean subgroup difference on the predictor approached one *SD*. However, when there was no difference on the criterion, the probability of detecting predictive against the reference group remained at or below chance level, irrespective of the mean subgroup difference on the predictor (see Table 4A and Figure 16).

Results of the focal group analysis showed that when there was no difference on the criterion, the rate of detecting predictive bias against the focal group increased from the chance level to approximately 55% as the mean subgroup difference on the predictor increased from zero to one *SD*. In contrast, when there was a mean subgroup difference on the criterion in favor of the reference group, detecting predictive bias against the focal group remained a chance level phenomenon, regardless of the mean subgroup differences on the predictor (see Table 4B and Figure 17). Hartigan and Wigdor (1989) found that only 3% of the studies had slope differences against the focal group, which is close to the chance-level findings in this study. Because the literature shows that there are mean subgroup differences on the criterion in favor of some

reference groups (e.g., males and Anglo-Americans), the chances of detecting predictive bias against some focal groups (e.g. females and African-Americans) are very low regardless of the mean subgroup difference on the predictor.

There is some statistical evidence to suggest that DIF may have a moderating effect on predictive bias. Results of the exploratory analysis showed that there was a significant three-way interaction among the predictor difference, criterion difference, and percentage of DIF variables in the detection of predictive bias against the focal group. The interaction can be explained in terms of DIF being present or not present.

When DIF was not present and there was a difference on the criterion, there was practically no relation between the subgroup differences on the predictor and the rate of detecting predictive bias against the focal group: It remained constant around 4%. However, when there was no difference on the criterion, predictive bias against the focal group increased as the mean difference on the predictor increased (Figure 18). When DIF was present in 30% of the items and there was a difference on the criterion, there was a slight increase in the detection of predictive bias against the focal group as the mean subgroup differences on the predictor increased, but the increase was small with the highest rate of detection being approximately 10%. In contrast, when there was no difference on the criterion, predictive bias against the focal group had a greater probability of being detected well above chance level (Figure 20). In other words, the evidence suggests that when DIF is present and there is a subgroup difference on the criterion, the odds of concluding that predictive bias against a focal group does not exist are rather high (i.e., from 49:1 to approximately 9:1). This means that researchers and practitioners are more likely to conclude that predictive bias does not exist against a focal group when a large amount of DIF is

present against a focal group. Thus, it is possible for DIF to be present while predictive bias is not detected. Before making any firm conclusions, however, the strengths and weakness of this study are highlighted.

*Strengths and weaknesses.* There are several strengths and weaknesses in this study. With respect to the strengths, most of the manipulations created conditions similar to what has been found in past research. The mean predictor difference variable was manipulated so that values would span the range of subgroup differences often reported in the literature, i.e., mean differences were from no difference to as high as one *SD* difference (Chung-Yan & Cronshaw, 2002; Hartigan & Wigdor, 1989). The validity coefficient variable was manipulated to reflect what has been reported in the literature about the validity of some cognitive ability tests (Hattrup & Schmitt, 1990; Muchinsky, 1993). The percentage of DIF and size of DIF variables were created to represent what has been observed in the research on IRT (Holland & Wainer, 1993). Moreover, the manipulation of the type of DIF (i.e., uniform) was done to reflect the kind of DIF often found in items (Hambleton et al., 1991; Swaminathan & Rogers, 1990). Also, the test was designed to have characteristics similar to the type of tests used in Industrial and Organizational Psychology (SIOP, 2003); it was intentionally created to have a large number of items and was dimensionally complex (Reckase, 1985). These strengths should be viewed along with the study's weaknesses before making claims about the generalizability of the findings.

This investigation also has several limitations. First, the sample size variable was manipulated to have equal subgroup proportions. Past research showed that differences in subgroup proportions commonly found in the literature adversely influence the power to detect differences in subgroup slopes (Stone-Romero et al., 1994). Thus, estimates of the rate of

detecting predictive bias in this study may be more liberal than what is found in samples with unequal subgroup proportions.

Second, all mean ability differences on the predictor and criterion were manipulated to be in favor of the reference group. The assumption of unidirectional mean score differences may not necessarily hold in all circumstances. However, subgroup differences on both the predictor and the criterion reported in the literature (Ford et al., 1986; Hough et al., 2001) are mainly unidirectional in favor of the reference group (e.g., Anglo-American) as compared to the focal group (e.g., African-American).

Third, DIF was manipulated in this study to reflect only uniform DIF. This may be viewed as a limitation in the sense that all possible types and directions of DIF, including DIF against both subgroups that cancel each other out, were not included. However, it may not be seen as a weakness because DIF of this type and direction allowed for a robust test of the relation between DIF and predictive bias under conditions that were most favorable to influence overall test scores and predictive bias, i.e., 30% of DIF, .9 effect size, and all DIF items were created to be in one direction against the focal group. To no avail, the results of this study showed unequivocally that predictive bias against the focal group was not influenced by DIF.

Fourth, some counterintuitive results were found for Hypothesis 4. For the reference group, as the mean difference on the predictor increased, the detection of predictive bias decreased. For the focal group, as the mean difference on the criterion increased, the detection of predictive bias decreased. These findings can be explained using Equations 12 and 13. Assume that there is an effects coded variable for subgroup membership (reference group = 1 and focal group = -1). From Equation 12, predictive bias is indicated by a significant coefficient,  $B_2$ .

Predictive bias against the reference group is present when  $B_2$  is positive and significant; predictive bias against the focal group is present when  $B_2$  is negative and significant. From Equation 13, three situations can occur: (a) when  $r_{yg} = (r_{yx})(r_{gx})$ ,  $B_2$  is zero and no predictive is present, (b) when  $r_{yg} > (r_{yx})(r_{gx})$ ,  $B_2$  is positive, thus increasing the chances of detecting predictive bias against the reference group, and (c) when  $r_{yg} < (r_{yx})(r_{gx})$ ,  $B_2$  is negative, thus increasing the chances of detecting predictive bias against the focal group.

In the present study, there was a negative relation between the mean predictor difference and predictive bias against the reference group. This is explained by the fact that as the predictor difference increases, the situation described in (c) is more likely to occur, thus increasing the chance of detecting predictive bias against the focal group because  $r_{gx}$  increases. However, the situation described in (b) is more likely to increase the chance of detecting predictive bias against the reference group when  $r_{gx}$  decreases. Therefore, predictive bias is more likely to occur against the reference group when the mean difference on the predictor is small as compared to when the mean difference on the predictor is large. The above reasoning is supported by the evidence of a positive relation between the predictor difference variable and the detection of predictive bias against the focal group and the negative relation between the predictor difference variable and predictive bias against the reference group. The same logic can be used to explain the counterintuitive results found in the focal group analysis.

There was a negative relation between the mean criterion difference variable and predictive bias in the focal group analysis. This can be explained by the fact that as the criterion difference increases under the conditions in this study, the situation described in (b) is more likely to occur, thus increasing the chance of detecting predictive bias against the reference

group because  $r_{yg}$  increases. However, the situation described in (c) is more likely to occur and increase the chance of detecting predictive bias against the focal group as  $r_{yg}$  approaches zero. Therefore, predictive bias is more likely to occur against the focal group when the mean difference on the criterion is zero as compared to when the mean difference on the criterion is greater than zero. This is also supported by the evidence of a positive relation between the criterion difference variable and the detection of predictive bias against the reference group and the negative relation between the criterion difference variable and predictive bias against the focal group. Given this explanation, future research should incorporate information about the relative contributions of  $r_{yg}$ ,  $r_{yx}$ , and  $r_{gx}$  before developing hypotheses about the relations among the predictor, criterion, and predictive bias for both focal and reference groups.

Finally, this study was a computer simulation. The generalizability of its findings is limited to the number of variables included in the study design. Predictor or criterion range restriction has been found to influence the rate of detecting predictive bias. In addition, the reliability of the predictor also has been reported to influence the rate of detecting predictive bias (Jensen, 1980). These variables were not included in the study. Thus, the generalizability to other places, settings, and times may suffer to the degree that the variables excluded from this study are operating in other settings. Notwithstanding these weaknesses, some conclusions can be advanced on the basis of the available evidence. These are considered next.

### *Conclusions*

In summary, there was no support for the notion that DIF has an influence on predictive bias against any subgroup. The evidence also suggests that predictive bias can exist against both

subgroups when DIF is not present (Table 4). Moreover, there was some support for the notion that when DIF is present and there is a mean subgroup difference on the criterion, it is likely that predictive bias against a focal group will not be detected. Evidence also supports the position that sample size influences the rate of predictive bias against both subgroups. Finally, the results also suggest that mean subgroup differences on the predictor and criterion interact in detecting predictive bias against both subgroups.

How do the results in the present study coincide with current notions about measurement bias and predictive bias? Hunter and Schmidt (Hunter & Schmidt, 2000; Hunter et al., 1984) speculated that if items are biased, then the test should be biased. It is interesting to note that Hunter and Schmidt (2000) provided no empirical evidence to support the premise that item bias leads to predictive bias. They reasoned that because the literature shows that cognitive ability tests are not biased in the predictive sense against African-Americans, therefore items are not biased against African-Americans. This conclusion rests upon the critical assumption that a direct relation exists between item bias (DIF) and predictive bias.

Because Hunter and Schmidt (2000) provided no evidence about the relation between item bias and predictive bias, the current study investigated this premise. The results suggest that items showing measurement bias (DIF) against a subgroup (e.g., African-Americans) have no influence on predictive bias against the same subgroup. So, the premise of the argument advanced by Hunter and Schmidt (2000) about DIF and predictive bias is not supported. The evidence in this study is consistent with the literature that attempts to deal with this question of the relation between measurement bias and predictive bias (Linn & Werts, 1971; Millsap, 1997, 1998). For example, Linn and Werts (1971) operationalized measurement bias as the lack of factorial invariance. The equality of subgroup latent intercepts, pattern matrices, and unique

factor variances is the definition of factorial invariance used in their study. Predictive bias was operationalized by subgroup differences in regression intercepts, slopes, or error variances (Cleary, 1968). They showed that predictive bias could still exist when there are no differences in subgroup factor structures. Using the same operational definitions of measurement bias and predictive bias as Linn and Werts (1971), Millsap (1997, 1998) proved mathematically that if factorial invariance holds for both the predictor and the criterion, then subgroup intercepts will differ when the common factor means differ. He concluded that a significant difference in subgroup intercepts has no implication for measurement bias in either the predictor or the criterion (Millsap, 1998). He also supported the view that measurement bias and predictive bias are not mutually supportive (Millsap, 1997).

Although the conclusions offered here are consistent with the results of other research, this investigation differs in several important respects. First, past studies operationalized measurement bias in the form of a common factor model and assumed that the relation between measurement bias and predictive bias was linear under a CTT framework. In the present study, however, measurement bias was assumed to be non-linear and was modeled using parametric IRT, which imposed some assumptions on the form of the ICC. Second, past studies employed small numerical examples (Drasgow, 1982; Linn & Werts, 1971) and mathematical proofs (Millsap, 1997, 1998). In this study, over 200,000 simulations of predictive bias were conducted under conditions commonly reported in the literature. Moreover, statistical hypothesis testing was employed, which was aided by illustrations and tables of the rate of detecting predictive bias against both reference and focal groups under different conditions. Overall, there is converging evidence that there is no relation between measurement bias and predictive bias.

Thus, based upon an unsupported premise about the relation between item bias and predictive bias, the inference of Hunter and Schmidt (Hunter & Schmidt, 2000; Hunter et al., 1984) is untenable given the evidence in this study. So, if future research also demonstrates the lack of a relation between measurement bias at the item level and predictive bias at the test score level, it may become clear that conclusions about the presence or absence of DIF cannot be made on the basis of evidence from a predictive bias analysis.

### *Scientific and Social Implications*

*Scientific implications.* There are several scientific implications of this research? First, this study provides some evidence that seriously questions the well-entrenched view that researchers and practitioners have about the mutually supportive relation between measurement bias and predictive bias. This study found evidence to support the view that predictive bias can exist when DIF is not present (Figure 18). The evidence in this research is also consistent with the notion that when DIF is present against a focal group, it is likely that predictive bias against the subgroup will not be detected. It should be noted that the conditions in this study are commonly present in both educational and employment contexts. Future research should consider the independent roles that studies of measurement bias at the item level and predictive bias at the test score level have in the test validation process.

Both predictive bias and DIF studies play key parts in test development and validation (AERA, APA, & NCME, 1999). They complement each other to the extent that each provides support for different aspects of validity, which is judged in an integrative fashion. In other words, the two types of bias studies are not substitutes for each other. Predictive bias studies provide

information in support of the relations that test scores have with other external variables. DIF studies are concerned with evidence pertaining to a consistent internal structure of the test for different subgroups. The recognition of these facts will perhaps lead to a better understanding of the test development and validation process.

Second, the central issue is not the influence of DIF items on subgroup observed score distributions but on the meaning and interpretation of scores from each subgroup. Because some researchers argue that DIF leads to small and insignificant mean score differences between subgroups, DIF has been relegated to an issue of no serious consequence. However, DIF has very important implications for inferences about the internal structure of the test (AERA, APA & NCME, 1999). From the traditional conception of validity (i.e., content, construct, and criterion-related), DIF is a phenomenon related to the construct validity of the test. Even if there were no differences in observed score subgroup distributions, can a consistent and meaningful interpretation of scores be given to members of different subgroups in the presence of DIF, assuming that an appropriate measurement model is used? If two subgroups have the same standing on one or more latent traits, then test items should be designed to measure these traits to the same degree in both subgroups so that questions about the internal structure of the test are not an issue. In other words, there should be equity in measurement. Drasgow (1984) refers to this as measurement equivalence. Thus, a consistent internal structure of the test is a necessary condition for an accurate interpretation of test scores, which provides a solid foundation for the analysis of other types of bias.

Third, although this study found no relation between measurement bias at the item level and predictive bias at the test score level, studies of the two types of bias have a proximal

relation to each other: Before researchers investigate the external relations that test scores have with external criteria, it is recommended that an evaluation of the evidence pertaining to the internal structure of the test be done. Other researchers advocate establishing measurement equivalence before conducting a predictive bias analysis (Hough et al., 2001). Similar to the effort given to selecting an unbiased criterion, equal attention should be given to the evaluation of the internal structure of the test and other aspects of validity. Whether it is through evaluating test content, conducting a factor analysis, reviewing substantive aspects of the construct, or investigating the internal structure through DIF analyses, designing tests that are fair and equitable to all should be one of the primary goals of measurement.

*Social implications.* There are some social implications of this research. As stated in the *Standards*, evidence pertaining to the social consequences of testing now plays an important role in forming an overall integrative judgment of validity. It has been noted that DIF has an influence on the internal structure of a test. To the extent that the construct is narrowly defined and other meaningful or irrelevant dimensions are measured, then DIF may occur (Roussos & Stout, 1996). Thus, a clear specification of the construct and the boundaries of the construct domain must be consistent with what is measured by the test. Sternberg (2000) argues that cultural dimensions play a critical role in determining what is judged as being important or intelligent. Because intelligence is often defined as general mental ability or ‘g,’ certain aspects of intelligence may not be represented in commonly developed tests of cognitive ability used in employment and educational settings in the United States (Sternberg, 2000; Sternberg & Hedlund, 2002). Due to the fact that most cognitive ability tests account for less than 30% of the variance in job performance and are known to cause adverse impact against minority groups, some researchers (e.g., Hough et al., 2001) advocate the measurement and use of motivational

factors, personality, and other aspects of intelligence that are not often considered in the development of a selection system (e.g., practical intelligence; Sternberg 2002).

The evidence in this study suggests that even when test items show DIF against a subgroup, the regression model is unlikely to detect predictive bias against the same subgroup. This raises a concern about the utility of the regression model as an indirect way to evaluate measurement bias. Other researchers and practitioners have also challenged the regression model on various grounds (Cascio, Outtz, Zedeck, & Goldstein, 1991; Chung-Yan & Cronshaw, 2002). For example, the use of the model in top-down selection has been known to cause adverse impact against minority groups (e.g., African-Americans and Hispanics) when the predictor is a cognitive ability test designed to measure crystallized intelligence (Hough et al., 2001; Sternberg & Hedlund, 2002). Cleary's model also fails to consider the excessively high false rejection rate of capable minority group members (Hartigan & Wigdor, 1989), which raises another point about the disproportionately negative outcomes for members of minority groups. It is the perception of disproportionately negative outcomes that will foster feelings of inequity (Adams, 1963). Moreover, organizations that desire workforce diversity and want to consider it in personnel selection decisions are not provided flexibility in meeting valued organizational objectives when cognitive ability testing and top-down selection are employed. Below is a promising alternative to the regression model when social and economic issues are taken into account.

Cascio et al. (2001) suggest the use of banding as an alternative to regression-based, top-down selection when cognitive ability tests are used. The standard error of the difference (SED) banding has been employed when the economic and social goals of the organization are

considered in a selection system. It is well known that all tests have a certain amount of error that some traditional selection methods do not account for in the selection process (Cascio et al., 1991; Zedeck, Outtz, Cascio, & Goldstein, 1991). Banding, which can be either fixed or sliding, considers the errors that are inherent in cognitive ability tests.

Most of the research on banding has evaluated the error in the predictor from a CTT perspective, i.e., the standard error of measurement is assumed to be constant and is used in computing the SED. Future research on banding could evaluate the degree to which error in the predictor from an IRT framework yields comparable selection results when sliding bands are used. IRT assumes that error varies as a function of  $\theta$  (see Figure 11); this assumption is more realistic than the assumption about the fixed error in a CTT framework (Embretson & Reise, 2000). Thus, more reasonable bands may be created by using the conditional standard error term from IRT.

Although SED banding has spawned some controversy (Schmidt, 1991), it has also inspired meaningful debate and has shown promise as a selection tool when both societal and economic goals of the organization are considered (Campion, Outtz, Zedeck, Schmidt, Kehoe, Murphy, & Guion, 2001). As the recent Supreme Court decision in the University of Michigan case supports the use of race as one of many factors in a narrowly tailored selection system, banding may be a feasible alternative by which organizations can achieve both workforce diversity and economic utility.

## APPENDIX A: PROOF OF THETA MAXIMUM AND ITEM INFORMATION

The following is a proof for theta maximum and the item information function (IIF) for a composite in a direction for the multidimensional 3-PL model. Formulas for item information and theta maximum ( $\theta_{\max}$ ) are well known for unidimensional 1-, 2-, and 3-PL models. However, IIFs and theta maximums are not well known for compensatory, multidimensional models (Reckase & McKinley, 1991; Segall, 1996). So, the purpose of this proof is to provide explicit formulas for item information in a direction and theta maximum for the Multidimensional 3-PL (M3-PL) model.

### *The Item Response Model*

The probability of a correct response for the M3-PL model (Reckase, 1997) is

$$P_i(\boldsymbol{\theta}_j) = c_i + (1 - c_i) [1 + \text{Exp}(-L)]^{-1}, \quad (\text{A1})$$

where  $L = D(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)$ ,  $D$  is equal to a scaling constant 1.7,  $\mathbf{a}_i$  is a vector of  $k$  discrimination parameters for item  $i$ ,  $[a_{1i}, a_{2i}, \dots, a_{ki}]'$ ,  $k$  is the number of dimensions,  $\boldsymbol{\theta}_j$  is a vector of  $k$  ability parameters for person  $j$ ,  $[\theta_{1j}, \theta_{2j}, \dots, \theta_{kj}]'$ , and  $d_i$  is a scalar related to difficulty. The probability of an incorrect response is given by  $Q_i(\boldsymbol{\theta}_j) = 1 - P_i(\boldsymbol{\theta}_j)$  or

$$Q_i(\boldsymbol{\theta}_j) = (1 - c_i) [1 + \text{Exp}(L)]^{-1}. \quad (\text{A2})$$

The point of steepest slope in the ability space is known as multidimensional discrimination,

$$MDISC_i = \|\mathbf{a}_i\| = (\mathbf{a}_i' \mathbf{a}_i)^{1/2}, \quad (\text{A3})$$

where  $\|\mathbf{a}_i\|$  represents the length of vector  $\mathbf{a}_i$  that is computed as the square root of the sum of squared elements of vector  $\mathbf{a}_i$ .  $MDISC_i$  is interpreted in the same manner as the discrimination parameter ( $a_i$ ) in unidimensional IRT. The difficulty of the item is the signed distance from the origin of the multidimensional space to the point of steepest slope. The formula for multidimensional difficulty is given by

$$MDIFF_i = -d_i(\|\mathbf{a}_i\|)^{-1} = -d_i/MDISC_i, \quad (A4)$$

and it is interpreted in the same way as the difficulty parameter ( $b_i$ ) in unidimensional IRT.

Reckase (1985) has shown  $MDIFF_i$  to be equal to the unidimensional measure of difficulty ( $b_i$ ) when there is only one dimension.

### *Item Information Function*

The IIF in a specific direction for the multidimensional logistic model (Reckase, 1997; Reckase & McKinley, 1991) is

$$I_{iu}(\boldsymbol{\theta}) = [\nabla P_i(\boldsymbol{\theta}_j) \cdot \mathbf{u}_i]^2 [P_i(\boldsymbol{\theta}_j)Q_i(\boldsymbol{\theta}_j)]^{-1}. \quad (A5)$$

$\nabla P_i(\boldsymbol{\theta}_j) \cdot \mathbf{u}_i$  is the directional derivative,  $\nabla P_i(\boldsymbol{\theta}_j)$  is the gradient. The vector of directional cosines,  $\mathbf{u}_i$ , is  $[a_{1i}/\|\mathbf{a}_i\|, a_{2i}/\|\mathbf{a}_i\|, \dots, a_{ki}/\|\mathbf{a}_i\|]'$  or  $[\cos \alpha_{1i}, \cos \alpha_{2i}, \dots, \cos \alpha_{ki}]'$ , where  $\cos \alpha_{ki}$  is the cosine of the angle ( $\alpha_{ki}$ ) from the axis orthogonal to dimension  $k$ . It should be noted that  $\|\mathbf{u}_i\| = 1$ . Next is the derivation of the IIF in a specified direction.

The first term in brackets within Equation A5 is the gradient of the function  $P_i(\boldsymbol{\theta}_j)$ , which can be written as

$$\nabla P_i(\boldsymbol{\theta}_j) = [\partial P_i(\boldsymbol{\theta}_j)/\partial \theta_{1j}, \partial P_i(\boldsymbol{\theta}_j)/\partial \theta_{2j}, \dots, \partial P_i(\boldsymbol{\theta}_j)/\partial \theta_{kj}]', \quad (A6)$$

where  $\partial P_i(\boldsymbol{\theta}_j)/\partial\theta_{kj}$  is the first partial derivative of  $P_i(\boldsymbol{\theta}_j)$  with respect to  $\theta_{kj}$ . The general

expression for the derivative of the  $k^{\text{th}}$  term in the gradient is as follows:

$$\begin{aligned} \partial P_i(\boldsymbol{\theta}_j)/\partial\theta_{kj} &= \partial\{c_i + (1 - c_i) [1 + \text{Exp}(-L)]^{-1}\}/\partial\theta_{kj} = -1(1 - c_i) [1 + \text{Exp}(-L)]^{-2} \\ \text{Exp}(-L) - Da_{ki} &= Da_{ki}(1 - c_i) [1 + \text{Exp}(L)]^{-1} [1 + \text{Exp}(-L)]^{-1}, \end{aligned} \quad (\text{A7})$$

or substituting Equation 2 into 7

$$\partial P_i(\boldsymbol{\theta}_j)/\partial\theta_{kj} = Da_{ki} Q_i(\boldsymbol{\theta}_j) [1 + \text{Exp}(-L)]^{-1}. \quad (\text{A8})$$

With the  $k$  elements of the gradient having the general form of Equation 8 and the elements of  $\mathbf{u}_i$

having the general form of  $a_{ki}/\|\mathbf{a}_i\|$ , the directional derivative of the function  $P_i(\boldsymbol{\theta}_j)$  in the

direction  $\mathbf{u}_i$  can be expressed as

$$\nabla P_i(\boldsymbol{\theta}_j) \cdot \mathbf{u}_i = D(\mathbf{a}'_i \mathbf{u}_i) Q_i(\boldsymbol{\theta}_j) [1 + \text{Exp}(-L)]^{-1}. \quad (\text{A9})$$

With the right-hand side of Equation A9 substituted in Equation A5, the IIF for the M3-PL model in a specified direction is

$$I_{iu}(\boldsymbol{\theta}) = \{D(\mathbf{a}'_i \mathbf{u}_i) Q_i(\boldsymbol{\theta}_j) [1 + \text{Exp}(-L)]^{-1}\}^2 [P_i(\boldsymbol{\theta}_j) Q_i(\boldsymbol{\theta}_j)]^{-1}. \quad (\text{A10})$$

or with some algebra, the IIF in a direction  $\mathbf{u}_i$  becomes

$$I_{iu}(\boldsymbol{\theta}) = D^2(\mathbf{a}'_i \mathbf{u}_i) Q_i(\boldsymbol{\theta}_j) \{P_i(\boldsymbol{\theta}_j) [1 + \text{Exp}(-L)]^{-1}\}^{-1}. \quad (\text{A11})$$

Corollary 1a. If it is assumed that there is no guessing (i.e.,  $c_i = 0$ ), the function in Equation A11

becomes the IIF in a direction for the M2-PL model,

$$I_{iu}(\boldsymbol{\theta}) = D^2(\mathbf{a}'_i \mathbf{u}_i) P_i(\boldsymbol{\theta}_j) Q_i(\boldsymbol{\theta}_j), \quad (\text{A12})$$

which is similar to the formula derived by Reckase and McKinley (1991).

Corollary 1b. If it is assumed that discrimination parameters on all of  $k$  dimensions are fixed at 1 and there is no guessing (i.e.,  $\mathbf{a}_i = [1, 1, \dots, 1]'$  and  $c_i = 0$ ), then Equation A11 reduces to the M1-PL,

$$I_{iu}(\boldsymbol{\theta}) = D^2 k P_i(\boldsymbol{\theta}_j) Q_i(\boldsymbol{\theta}_j), \quad (\text{A13})$$

where  $k$  is equal to the number of dimensions.

Corollary 1c. If there is only one dimension, then Equations A13, A12, and A11 become the item information functions for the unidimensional 1-, 2-, and 3-PL models, respectively.

#### *Theta Maximum for the Multidimensional 3-PL Model*

The formula for the location of maximum item information or theta maximum in a specified direction is derived by setting the directional derivative of the IIF for the M3-PL model equal to zero,

$$\nabla I_{iu}(\boldsymbol{\theta}) \cdot \mathbf{u}_i = 0, \quad (\text{A14})$$

where  $\nabla I_{iu}(\boldsymbol{\theta})$  is  $[\partial I_{iu}(\boldsymbol{\theta})/\partial\theta_{1j}, \partial I_{iu}(\boldsymbol{\theta})/\partial\theta_{2j}, \dots, \partial I_{iu}(\boldsymbol{\theta})/\partial\theta_{kj}]'$  and  $\mathbf{u}_i$  was defined earlier. The IFF in Equation A11 is expressed in a different form as

$$I_{iu}(\boldsymbol{\theta}) = D^2 (\mathbf{a}_i' \mathbf{u}_i)^2 (1 - c_i) \text{Exp}(-L) \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-1}, \quad (\text{A15})$$

where  $L$  is the logit, which is equal to  $D(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)$ . The general form of the  $k^{\text{th}}$  element of the gradient,  $\nabla I_{iu}(\boldsymbol{\theta})$ , is derived using the quotient, product, and chain rules of calculus.

$$\begin{aligned} \partial I_{iu}(\boldsymbol{\theta}) / \partial\theta_{kj} &= \partial D^2 (\mathbf{a}_i' \mathbf{u}_i)^2 (1 - c_i) \text{Exp}(-L) \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-1} / \partial\theta_{kj} \\ &= \{-D a_{ki} D^2 (\mathbf{a}_i' \mathbf{u}_i)^2 (1 - c_i) \text{Exp}(-L) [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 - \end{aligned}$$

$$\begin{aligned}
& \{2[1 + \text{Exp}(-L)] \text{Exp}(-L) - Da_{ki}[1 + c_i \text{Exp}(-L)] + [1 + \text{Exp}(-L)]^2 c_i \text{Exp}(-L) - Da_{ki}\} D^2 (\mathbf{a}'_i \mathbf{u}_i)^2 (1 - \\
& c_i \text{Exp}(-L)) \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-2} \\
& = \{-a_{ki} D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 + \\
& a_{ki} 2 \text{Exp}(-L) [1 + \text{Exp}(-L)] [1 + c_i \text{Exp}(-L)] D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] + \\
& a_{ki} c_i \text{Exp}(-L) [1 + \text{Exp}(-L)]^2 D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-2} \\
& = \{a_{ki} D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \{-1 - c_i \text{Exp}(-L) + c_i \text{Exp}(-L)\} + \\
& a_{ki} 2 \text{Exp}(-L) [1 + \text{Exp}(-L)] [1 + c_i \text{Exp}(-L)] D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)] [1 + \\
& \text{Exp}(-L)]^2 \}^{-2} \\
& = \{a_{ki} D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 (-1) + a_{ki} 2 \text{Exp}(-L) [1 + \text{Exp}(-L)] \\
& [1 + c_i \text{Exp}(-L)] D^3 (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-2} \\
& = D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)] \{ (-1) [1 + \text{Exp}(-L)] + 2 \text{Exp}(-L) \\
& [1 + c_i \text{Exp}(-L)] \} \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-2} \\
& = D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)] \{ (-1) [1 + \text{Exp}(-L)] + 2 \text{Exp}(-L) \\
& [1 + c_i \text{Exp}(-L)] \} \{ [1 + c_i \text{Exp}(-L)]^2 [1 + \text{Exp}(-L)]^4 \}^{-1} \\
& = D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ (-1) [1 + \text{Exp}(-L)] + 2 \text{Exp}(-L) \\
& [1 + c_i \text{Exp}(-L)] \} \{ [1 + c_i \text{Exp}(-L)]^2 [1 + \text{Exp}(-L)]^3 \}^{-1} \\
& = -D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)]^2 [1 + \text{Exp}(-L)]^3 \}^{-1} + \\
& D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] 2 \text{Exp}(-L) [1 + c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)]^2 [1 + \text{Exp}(-L)]^3 \}^{-1}
\end{aligned}$$

$$\begin{aligned}
&= -D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)]^2 [1 + \text{Exp}(-L)]^2 \}^{-1} + \\
D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 &[\text{Exp}(-L) - c_i \text{Exp}(-L)] 2\text{Exp}(-L) \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^3 \}^{-1}. \tag{A16}
\end{aligned}$$

From Equation A16, the  $k^{\text{th}}$  element of the gradient  $\nabla I_{iu}(\boldsymbol{\theta})$  can be expressed as

$$\begin{aligned}
\partial I_{iu}(\boldsymbol{\theta}) / \partial \theta_{kj} &= [D^3 a_{ki} (\mathbf{a}'_i \mathbf{u}_i)^2 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-1}] \\
&\{-1[1 + c_i \text{Exp}(-L)]^{-1} + 2\text{Exp}(-L)[1 + \text{Exp}(-L)]^{-1}\}. \tag{A17}
\end{aligned}$$

With the  $k$  elements of the gradient,  $\nabla I_{iu}(\boldsymbol{\theta})$ , having the general form of Equation A17 and the elements of  $\mathbf{u}_i$  have the general form of  $a_{ki}/\|\mathbf{a}_i\|$ , the directional derivative of the IIF,  $I_{iu}(\boldsymbol{\theta})$ , in the direction  $\mathbf{u}_i$  is expressed as

$$\begin{aligned}
\nabla I_{iu}(\boldsymbol{\theta}) \cdot \mathbf{u}_i &= [D^3 (\mathbf{a}'_i \mathbf{u}_i)^3 [\text{Exp}(-L) - c_i \text{Exp}(-L)] \{ [1 + c_i \text{Exp}(-L)] [1 + \text{Exp}(-L)]^2 \}^{-1}] \\
&\{-1[1 + c_i \text{Exp}(-L)]^{-1} + 2\text{Exp}(-L)[1 + \text{Exp}(-L)]^{-1}\} = 0. \tag{A18}
\end{aligned}$$

From Equation A18, item information is maximized when the following condition is satisfied:

$$-1[1 + c_i \text{Exp}(-L)]^{-1} + 2\text{Exp}(-L)[1 + \text{Exp}(-L)]^{-1} = 0, \tag{A19}$$

or

$$P_i(\boldsymbol{\theta}) = .5\text{Exp}(L). \tag{A20}$$

Equation A20 is a necessary condition for maximizing item information for the unidimensional and multidimensional 1-, 2-, and 3-PL models. In the unidimensional 1-PL model, information is maximized when  $P_i(\boldsymbol{\theta}) = Q_i(\boldsymbol{\theta}) = .5$ . When  $P_i(\boldsymbol{\theta})$  is equal to .5, then the natural logarithm of the odds of getting the item correct is 0, i.e.,  $\ln[P_i(\boldsymbol{\theta})/Q_i(\boldsymbol{\theta})] = 0$ , which implies that  $L$  in (20) is 0 and  $\text{Exp}(L) = 1$ . This leaves the well-known condition for maximizing information in the 1- and 2-PL cases, which is  $P_i(\boldsymbol{\theta}) = Q_i(\boldsymbol{\theta}) = .5$ . In classical test theory, this is

akin to maximizing item variance at  $p_i = .5$ . For the 3-PL unidimensional and multidimensional models, the right-hand side of Equation A20 and the probability of a correct response are not equal to .5 when information is maximized, but the equality is satisfied at a different value, which is primarily a function of guessing. Equation A20 can also be written as

$$2\text{Exp}(-L) = [P_i(\boldsymbol{\theta})]^{-1}. \quad (\text{A21})$$

The probability of a correct response can also be written as

$$P_i(\boldsymbol{\theta}) = [\text{Exp}(L) + c_i][\text{Exp}(L) + 1]^{-1}. \quad (\text{A22})$$

With the right-hand side of Equation A22 substituted into Equation A21,

$$2\text{Exp}(-L) = [\text{Exp}(L) + 1][\text{Exp}(L) + c_i]^{-1}. \quad (\text{A23})$$

To solve for  $L$ , Equation A23 is written as

$$2 + 2c_i\text{Exp}(-L) = \text{Exp}(L) + 1. \quad (\text{A24})$$

After multiplying both sides by  $\text{Exp}(L)$ ,

$$2\text{Exp}(L) + 2c_i = \text{Exp}(2L) + \text{Exp}(L), \quad (\text{A25})$$

which, after subtracting  $2\text{Exp}(L)$  from both sides, is equivalent to

$$2c_i = \text{Exp}(2L) - \text{Exp}(L). \quad (\text{A26})$$

After multiplying both sides by 4, Equation A26 becomes

$$8c_i = 4 \text{Exp}(2L) - 4\text{Exp}(L). \quad (\text{A27})$$

When 1 is added to both sides, then Equation A27 becomes

$$8c_i + 1 = 4 \text{Exp}(2L) - 4\text{Exp}(L) + 1. \quad (\text{A28})$$

By way of the binomial theorem, the equality of Equation A28 can be written as

$$8c_i + 1 = [2 \text{Exp}(L) - 1]^2, \quad (\text{A29})$$

which after taking the square root of both sides becomes

$$(8c_i + 1)^{1/2} = 2 \text{Exp}(L) - 1. \quad (\text{A30})$$

Solving for  $L$  in Equation A30 gives

$$L = \ln\{.5 [1 + (8c_i + 1)^{1/2}]\}, \quad (\text{A31})$$

which becomes

$$D(\mathbf{a}'_i \boldsymbol{\theta} + d_i) = \ln\{.5 [1 + (8c_i + 1)^{1/2}]\}. \quad (\text{A32})$$

Now the objective is to solve for the vector  $\boldsymbol{\theta}$  that maximizes the information function of the M3PL model in the direction  $\mathbf{u}_i$ . To that end,

$$\mathbf{a}'_i \boldsymbol{\theta} = \ln\{.5 [1 + (8c_i + 1)^{1/2}]\} D^{-1} - d_i. \quad (\text{A33})$$

Because  $\mathbf{u}'_i = \mathbf{a}'_i / \|\mathbf{a}_i\|$ , both sides are divided by  $\|\mathbf{a}_i\|$  or  $MDISC_i$  that results in

$$\mathbf{u}'_i \boldsymbol{\theta} = \ln\{.5 [1 + (8c_i + 1)^{1/2}]\} (D\|\mathbf{a}_i\|)^{-1} - d_i (\|\mathbf{a}_i\|)^{-1}, \quad (\text{A34})$$

Now both sides are pre-multiplied by  $\mathbf{u}_i$ ,

$$\mathbf{u}_i \mathbf{u}'_i \boldsymbol{\theta} = \mathbf{u}_i [\ln\{.5 [1 + (8c_i + 1)^{1/2}]\} (D\|\mathbf{a}_i\|)^{-1} - d_i (\|\mathbf{a}_i\|)^{-1}]. \quad (\text{A35})$$

Because the associative law holds for multiplication of matrices, the left-hand side of Equation A35 can be written as  $(\mathbf{u}_i \mathbf{u}'_i) \boldsymbol{\theta}$ . The product of  $(\mathbf{u}_i \mathbf{u}'_i)$  is a  $k$  by  $k$  matrix,  $\mathbf{U}$ , which has a few properties that should be noted: (1) It is symmetric, thus  $\mathbf{U} = \mathbf{U}'$ , (2) The determinant of the matrix  $\mathbf{U}$  is zero, thus it is singular, (3)  $\mathbf{U}^2 = \mathbf{U}' \mathbf{U} = \mathbf{U} \mathbf{U} = \mathbf{U}$ , thus  $\mathbf{U}$  is idempotent, (4) The main diagonal elements of  $\mathbf{U}$  are  $\cos^2 \alpha_{ki}$ , thus the trace of  $\mathbf{U}$ , i.e.,  $\text{tr}(\mathbf{U})$ , is equal to one, and (5) The rank of an idempotent matrix is equal to its trace, thus the rank of  $\mathbf{U}$  is equal to one. With  $\mathbf{U}$  substituted for  $\mathbf{u}_i \mathbf{u}'_i$ , Equation A35 is written as

$$\mathbf{U}\boldsymbol{\theta} = \mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}]. \quad (\text{A36})$$

Pre-multiply both sides by  $\mathbf{U}$  yields

$$\mathbf{U}\mathbf{U}\boldsymbol{\theta} = \mathbf{U}\mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}]. \quad (\text{A37})$$

By the third property mentioned above for  $\mathbf{U}$ , Equation A37 can be written as

$$\mathbf{U}\boldsymbol{\theta} = \mathbf{U}\mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}]. \quad (\text{A38})$$

Equation A38 is written as

$$\mathbf{U}\boldsymbol{\theta} - \mathbf{U}\mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}] = \mathbf{0}, \quad (\text{A39})$$

where  $\mathbf{0}$  is the same size as  $\mathbf{u}_i$  or by the distributive law for matrices,

$$\mathbf{U}\{\boldsymbol{\theta} - \mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}]\} = \mathbf{0}. \quad (\text{A40})$$

Solutions that satisfy Equation A40 are when  $\mathbf{U}$  is equal to a null matrix ( $\mathbf{O}$ ) and when

$$\boldsymbol{\theta} = \mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}]. \quad (\text{A41})$$

Therefore, theta maximum in a specified direction for the M3-PL model is

$$\boldsymbol{\theta}_{\max} = \mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D\|\mathbf{a}_i\|)^{-1} - d_i(\|\mathbf{a}_i\|)^{-1}], \quad (\text{A42})$$

or substituting values of Equations A3 and A4 into Equation A42

$$\boldsymbol{\theta}_{\max} = \mathbf{u}_i[\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D \cdot MDISC_i)^{-1} + MDIFF_i]. \quad (\text{A43})$$

Several corollaries follow from this result in Equation A43.

Corollary 2a. The location on the  $k^{\text{th}}$  dimension where information is maximized is given by

$$\boldsymbol{\theta}_{\max k} = [\ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(D \cdot MDISC_i)^{-1} + MDIFF_i] \cos \alpha_{ki}. \quad (\text{A44})$$

Corollary 2b. If it is assumed that there is no guessing, i.e.,  $c_i = 0$ , Equation A43 reduces to

$$\theta_{\max} = MDIFF_i u_i. \quad (A45)$$

Equation A45 is the same as that implied by Reckase and McKinley (1991) for the location of maximum item information for the multidimensional 2-PL model.

Corollary 2c. If it is assumed that there is only one dimension, then Equation A43 reduces to the well-known formula for theta maximum derived by Birnbaum (1968),

$$\theta_{\max} = \ln\{.5 [1 + (8c_i + 1)^{1/2}]\}(Da_i)^{-1} + b_i. \quad (A46)$$

IIFs and formulas for the location of maximum item information for the multidimensional 1-, 2-, and 3-PL models are listed in Table 2.

## APPENDIX B: ALGORITHM FOR THE SIMULATION STUDY

The following is a description of the algorithm that was used in this study. The sub-routines are **1A**, **1B**, **2**, and **3**. This routine produced 500 replications of each condition in this study. See Figure 21 for a diagram of the program architecture.

*1A. Create Subgroup Distributions ( $D_{nx3}$ ), Manipulate Ability Differences and Validity Coefficients.*

This is represented by the Subgroup Object in Figure 21.

$D_{nx3}$  consists of ability matrix ( $A_{nx2}$ ) and criterion vector ( $y_{nx1}$ ).

1. Create Multivariate Normal Distributions. Specify:
  - (a) Sample Sizes (35, 70, 105)
  - (b) Mean vector ( $\mu_{3x1}$ ) according to predictor difference and criterion difference manipulations.
  - (c) Covariance matrix ( $\Sigma_{3x3}$ )
    - (1) validity coefficient manipulation is specified here.
2. Generate  $D_{nx3}$  for the reference group ( $R_{nx3}$ ) and focal group ( $F_{nx3}$ ) so that mean difference in composite distributions can be manipulated. (0, .33, .66, 1)
3. Store  $F_{nx3}$  and  $R_{nx3}$  in an array,  $C_{nx3x2}$
4. Store 48 unique conditions of  $C$  in a dictionary.

*1B. Create Item Parameters and Manipulate DIF Conditions.*

This is represented by the Item and Test Objects in Figure 21.

1. Generate  $d_i$ ,  $a_i$  according to best measurement direction  $u$  for base test of 60 items. See Table 3.

2. Manipulate percentage of DIF: 0, 15%, and 30%. See Method section.
3. Manipulate effect size of DIF for DIF items: .3, .6, and .9. See Methods section.

The percentage and effect size of DIF conditions will be crossed so that a total of nine (9) sets of 60 test items will be created.

4. For each test form, store each set of item parameters in  $T_{60 \times 3}$
5. Store 9 unique test sets of  $T_{60 \times 3}$  in a dictionary.

## 2. Simulate Examinees Taking Test.

This is represented by the Condition and Item Objects in Figure 21.

- 1A. Retrieve subgroup distribution ( $F_{nx3}$  or  $R_{nx3}$ ) from  $C_{nx3 \times 2}$  in dictionary of *IA*.
- 1B. Retrieve a set of test items ( $T_{60 \times 3}$ ) from dictionary of *IB*.
2. Partition  $D_{nx3}$  so that only the ability matrix ( $A_{nx2}$ ) is present. Store criterion vector ( $y_{nx1}$ ), which will be attached to simulated total test score vector ( $s_{nx1}$ ).
3. Expose an examinee in  $D_{nx3}$  to every item in the test ( $T_{60 \times 3}$ ) according to manipulation.
4. Use M2-PL model to produce each probability. An examinee's exposure to an item will produce a probability; there will be 60 item probabilities for each examinee.
5. Compare each item probability to a randomly generated number from a uniform distribution (0,1). If the probability is greater than or equal to the randomly generated number, the simulated dichotomous response is 1 for correct; otherwise, 0 for incorrect. Store vector of each examinee's responses as a row (1x60).
6. Repeat 3-5 for each of  $N$  examinees in subgroup.
7. Bind all rows of response from (5) to produce a matrix,  $Q_{nx60}$ , of dichotomous responses to test questions.

8. For each examinee, sum the 60 dichotomous responses to produce a total test score. Repeat for each examinee and store the  $n$  test scores in a test score vector,  $\mathbf{s}_{nx1}$ .
9. Recall criterion score vector ( $\mathbf{y}_{nx1}$ ) and bind to  $\mathbf{s}_{nx1}$  to produce matrix,  $\mathbf{S}_{nx2}$ .
10. Create a vector ( $\mathbf{g}_{nx1}$ ) to represent subgroup membership.
11. Bind  $\mathbf{g}$  vector to  $\mathbf{S}_{nx2}$  to create a  $\mathbf{M}_{nx3}$  matrix for the subgroup.
12. Repeat algorithm for each subgroup.
13. Bind the  $\mathbf{M}_{nx3}$  for each subgroup to produce overall matrix  $\mathbf{M}_{Nx3}$  with reference and focal group members representing rows and the criterion, subgroup membership, and test scores representing columns.

### 3. Set-Up Data and Conduct Predictive Bias Analyses

This is represented by the Regression, Condition, and Study Objects in Figure 21.

1. For each  $\mathbf{M}_{Nx3}$  in (2), partition  $\mathbf{M}$  to produce a matrix  $\mathbf{X}_{Nx2}$ , where the first column is subgroup membership and the second column is the test score. There will also be a vector of criterion scores ( $\mathbf{y}_{Nx1}$ ).
2. Center scores in  $\mathbf{X}_{Nx2}$  such that the mean of subgroup membership is zero (0) and the mean of test scores is zero (0).
3. Create interaction vector  $\mathbf{x}_{Nx1}$  by multiplying within each row the centered subgroup membership variable and centered test score.
4. Bind the interaction vector  $\mathbf{x}_{Nx1}$  to the centered matrix of  $\mathbf{X}_{Nx2}$  to produce a new  $\mathbf{X}_{Nx3}$  of centered scores with columns representing subgroup membership, test scores (predictor) and the test score by subgroup membership interaction.

5. Bind a constant vector of ones (1)  $\mathbf{o}_{N \times 1}$  to the result,  $\mathbf{X}_{N \times 3}$ , in 4 to produce the matrix of predictors ( $\mathbf{X}_{N \times 4}$ ). This is the matrix represented in Equations 2, 3, and 5.
6. Conduct predictive bias analysis consistent with Equation 2, using  $\mathbf{y}_{N \times 1}$  as the criterion.
7. Code results as either detecting predictive bias or not detecting predictive bias.
8. Store result in a vector,  $\mathbf{w}$ .
9. Loop 500 times for each combination of values in the dictionaries produced in *IA* and *IB*.

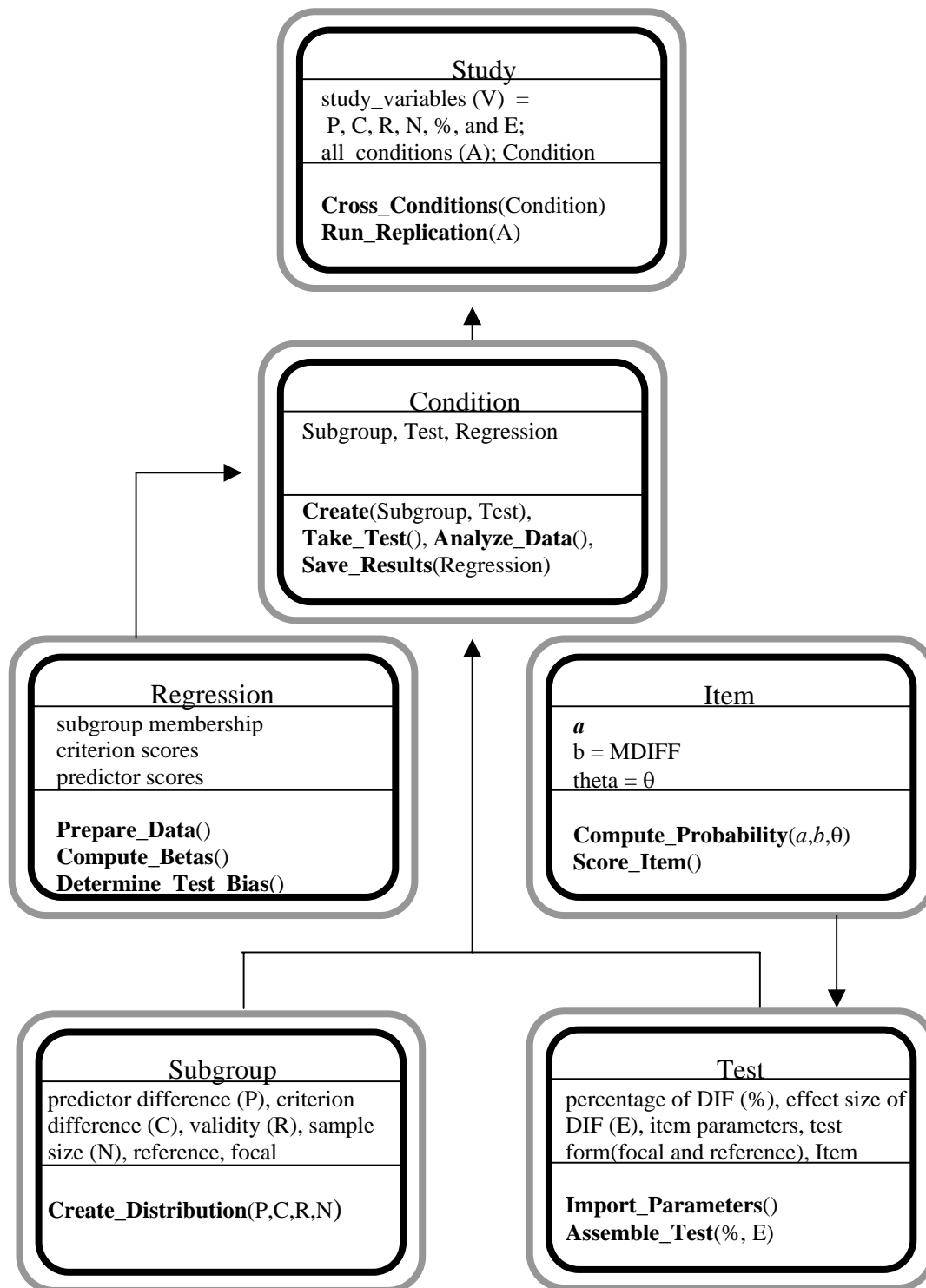


Figure 21. Diagram of the object-oriented program architecture.

## REFERENCES

- Adams, J. S. (1963). Toward and understanding of inequity. *Journal of Abnormal and Social Psychology, 67*, 422-436.
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Ackerman, T. A. (1994a). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278.
- Ackerman, T. A. (1994b). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18*, 257-275.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications, Inc.
- Aiken, L. S., West, S. G., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality, 64*, 1-48.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.
- American Psychological Association (1974). *Standards for educational and psychological tests*. Washington, DC: Author.

- American Psychological Association, Division of Industrial and Organizational Psychology.  
(1980). *Principles for the validation and used of personnel selection procedures* (2<sup>nd</sup> ed.).  
Washington, DC: Author.
- Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative in differential analysis. *Personnel Psychology, 31*, 233-241.
- Bartlett, C. J., Bobko, P., & Pine, S. M. (1977). Single-group validity: Fallacy of the facts? *Journal of Applied Psychology, 62*, 155-157.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores* (pp. 453-479). Reading, MA: Addison-Wesley.
- Boehm, V. R. (1972). Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology, 56*, 33-39.
- Bryant, D. U. (in press). A note on item information in any direction for the three-parameter logistic multidimensional model. *Psychometrika*.
- Bryant, D. U., Williamson, D., Wooten, W., & Forde, D. S. (2004, April). *Differential item functioning and item information*. Paper presented at the meeting of the Society for Industrial and Organizational Psychology. Chicago, IL.
- Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*, (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications, Inc.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.
- Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy in score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology*, *54*, 149-185.
- Cascio, W. F. (1998). *Applied psychology in human resource management* (5<sup>th</sup> ed). Upper Saddle River, NJ: Prentice Hall.
- Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, *4*, 233-264.
- Clauser, B. E., Nungester, R. J., Mazor, K. M., & Ripley, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement*, *33*, 202-214.
- Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, *33*, 453-464.
- Cleary, T. A. (1968). Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115-124.
- Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, *28*, 61-75.

- Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1978). Educational uses of tests with disadvantaged students. *American Psychologist, 30*, 15-41.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlational analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Erlbaum.
- Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10*, 237-255.
- Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3<sup>rd</sup> ed., pp. 201-219). New York, NY: Macmillan.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.
- Chung-Yan, G. A., & Cronshaw, S. F. (2002). A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. *Journal of Occupational and Organizational Psychology, 75*, 489-509.
- Darlington, R. B. (1990). *Regression and linear models*. New York, NY: McGraw-Hill, Inc.
- Dimitrov, D. M. (2003). Marginal true score measures and reliability of binary item as a function of their IRT parameters. *Applied Psychological Measurement, 27*, 440-458.
- Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.
- Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92*, 526-531.

- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, 95, 134-135.
- Ellis, B. B., & Mead, A. D. (2002). Item analysis: Theory and practice using classical and modern test theory. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 324-343). Malden, MA: Blackwell Publishing, Inc.
- Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Embretson, S. E. (1995). Developments toward a cognitive design system for psychological tests. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 17-48). Palo Alto, CA: Davies-Black Publishing.
- Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Lawrence Erlbaum Associates.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Equal Employment Opportunity Commission (1970). Guidelines on employee selection procedures. *Code of federal regulations* (Title 29, Chapter XIV, Section 1607). Washington, DC: US Government Printing Office.

- Florida Department of Education (2002). *Technical report: For operational administrations of the Florida comprehensive assessment test 2000*. Tallahassee, FL: Harcourt Educational Measurement.
- Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin, 99*, 330-337.
- Gierl, M. J., Henderson, D., Jodoin, M., & Klinger, D. (2001). Minimizing the influence of item parameter estimation errors in test development: A comparison of three selection procedures. *The Journal of Experimental Education, 69*, 261-279.
- Gulliksen, H. (1950). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Hambleton, R. K., & de Gruiter, D. N. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement, 20*, 355-367.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park CA: Sage.
- Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the general aptitude test battery*. Washington, DC: National Academy Press.
- Hattrup, K., & Schmitt, N. (1990). Prediction of trades apprentices' performance on job sample criteria. *Personnel Psychology, 43*, 453-466.

- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2<sup>nd</sup> ed.). New York: John Wiley & Sons, Inc.
- Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issue, evidence and lessons learned. *International Journal of Selection and Assessment*, 9, 152-194.
- Humphreys, L. G. (1973). Statistical definitions of test validity for minority groups. *Journal of Applied Psychology*, 58, 1-4.
- Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, 6, 151-158.
- Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds and R. T. Brown (Eds.), *Perspectives in mental testing* (pp. 41-100). New York, NY: Plenum Press.
- Huysamen, G. K. (2002). The relevance of the new APA standards for educational and psychological test for employment testing in South Africa. *South African Journal of Psychology*, 32, 26-33.
- Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.
- Jones, L. V., & Applebaum, M. I. (1989). Psychometric methods. *Annual Review of Psychology*, 40, 23-43.

- Junker, B. J. (1992, April). *Ability estimation in unidimensional models when more than one trait is present*. Paper presented at the meeting of the American Educational Research Association. San Francisco, CA.
- Kyllonen, P. C., & Christal, R. E. (1989). Cognitive modeling of learning abilities: A status report of LAMP. In R. F. Dillion, & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied issues*. (pp. 146-173). New York, NY: Praeger Publishing.
- Lewis-Beck, M. S. (1980). *Applied regression: An introduction*. Beverly Hills, CA: Sage Publications.
- Linn, R. J. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, *63*, 507-512.
- Linn, R. J., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, *8*, 1-4.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lutz, M., & Ascher, D. (1999). *Learning python*. Sebastopol, CA: O'Reilly and Associates, Inc.
- Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of noharm estimates. *Journal of Educational and Behavioral Statistics*, *26*, 51-71.

- Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.
- McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning*, (pp. 389-396). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.
- Mellenbergh, G. J. (1982). Contingency table models for assessing bias. *Journal of Educational Statistics, 7*, 105-118.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3<sup>rd</sup> ed., pp. 13-103). New York, NY: Macmillan.
- Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.
- Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

- Messick, S. (1996). Human abilities and modes of attention: The issue of stylistic consistencies in cognition. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 77-96). Mahwah, NJ: Lawrence Erlbaum Associates.
- Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education, 5*, 193-211.
- Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248-260.
- Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*, 403-424.
- Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics: Third edition*. New York, NY: McGraw-Hill.
- Muchinsky, P. (1993). Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs. *Journal of Business and Psychology, 7*, 373-382.
- Nunnally, J. C., & Bernstein, I. C. (1994). *Psychometric theory* (3<sup>rd</sup> ed.). New York, NY: McGraw-Hill, Inc.
- Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*, 237-248.
- Osterlind, S. J. (1983). *Test item bias*. Beverly Hills, CA: Sage Publication, Inc.
- Peduzzi, P. N., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 99*, 1373-1379.

- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197-207.
- Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement*, *53*, 301-314.
- Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 156-188). San Francisco, CA: Jossey-Bass.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401-412.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361-373.
- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer-Verlag.
- Rencher, A. C. (2000). *Linear models in statistics*. New York, NY: John Wiley & Sons, Inc.
- Roussos, L., & Stout, W. (1996). DIF from the multidimensional perspective. *Applied Psychological Measurement*, *20*, 335-371.
- Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, *5*, 213-233.

- Sackett, P. R., Schmitt, N., Ellingson, J. E., Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56*, 302-318.
- Schmidt, F. L. (1991). Why banding procedures in personnel selection are logically flawed. *Human Performance, 4*, 265-277.
- Schmidt, F. L., & Hunter, J. E. (1974). Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist, 29*, 1-8.
- Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 11, pp. 115-139). Chichester, UK: Wiley.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.
- Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures: Fourth edition*. Bowling Green, OH: Society for Industrial and Organizational Psychology.
- Sternberg, R. J. (2000). Implicit theories of intelligence as exemplar stories of success: Why intelligence test validity is in the eye of the beholder. *Psychology, Public Policy, and Law, 6*, 159-167.
- Sternberg, R. J., & Hedlund, J. (2002). Practical intelligence, g, and work psychology. *Human Performance, 15*, 143-160.

- Stone, E. F. (1986). Research methods in industrial and organizational psychology: Selected issues and trends. In C. L. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 305-334). New York, NY: Wiley.
- Stone, E. F. (1988). Moderator variables in research: A review and analysis of conceptual and methodological issues. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management* (Vol 6, pp. 191-229). Greenwich, CT: JAI Press.
- Stone, E. F., & Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology*, *74*, 3-10.
- Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderated effects of dichotomous variables. *Journal of Management*, *1*, 167-178.
- Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology*, *79*, 354-359.
- Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensional assessment. *Applied Psychological Measurement*, *20*, 331-354.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.
- Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement*, *8*, 63-70.

Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computer adaptive testing. In D. J. Lubinski and R. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings*, (pp. 49-79). Palo Alto, CA: Davies-Black Publishing.

Zedeck, S., Outtz, J., Cascio, W. F., & Goldstein, I. L. (1991). Why do “testing experts” have such limited vision? *Human Performance*, 4, 297-308.